Extension des méthodes d'analyse factorielles à des données de type intervalle

Ahlame Douzal-Chouakria

THÈSE

présentée à

l'Université Paris IX–Dauphine
U.F.R. Sciences des organisations

pour l'obtention du titre de

DOCTEUR EN SCIENCES

Spécialité

INFORMATIQUE
(arrêté du 30 mars 1992)

par

Ahlame CHOUAKRIA

Sujet de la thèse :

Extension des méthodes d'analyse factorielle à des données de type intervalle

Soutenue le 10 septembre 1998 devant le jury composé de :

MM.	**Ludovic LEBART**	Président et Suffragant
	Pierre CAZES	Directeurs
	Edwin DIDAY	
	Gérard GOVAERT	Rapporteurs
	Gilbert SAPORTA	

“L’université n’entend donner aucune approbation ni improbation aux opinions émises dans les thèses : ces opinions doivent être considérées comme propres à leurs auteurs.”

RESUME

L'analyse factorielle permet d'extraire, à partir de données nombreuses, les tendances les plus marquantes. À l'aide de représentations graphiques, elle visualise des groupements, des oppositions, des tendances, impossibles à discerner directement sur un tableau de données. L'objectif de ce travail est d'étendre, deux méthodes en analyse factorielle, l'analyse en composantes principales et l'analyse des correspondances multiples, à des données de type intervalle. Dans la première partie de ce travail, on présente une extension de l'analyse en composantes principales à des données de type intervalle, munies éventuellement de contraintes de domaines. On propose deux nouvelles approches : la méthode des sommets et la méthode des centres. L'inertie prise en compte par chacune de ces méthodes est analysée puis comparée. Les paramètres d'aide à l'interprétation sont généralisés aux données intervalles. La visualisation ponctuelle des individus, dans les plans factoriels, se traduit ici par une visualisation de rectangles, segments ou points. On propose une procédure itérative qui fournit une visualisation des objets à différents niveaux de qualité de représentation. D'une part, on confronte dans un cadre probabiliste la méthode des sommets et la méthode des centres. D'autre part, on établit un rapprochement entre la méthode des sommets, la méthode STATIS et l'analyse factorielle discriminante. Dans une deuxième partie, on s'intéresse à l'extension de l'analyse des correspondances multiples à des données intervalles. On propose trois techniques de codage de variables de type intervalle : le codage croisé, le codage par sommets et le codage sans décomposition. Les deux premières techniques se basent sur la décomposition des variables intervalles en variables numériques. La dernière technique se base sur l'extension d'outils de codage des variables numériques (fonction de répartition, histogramme, fonction d'appartenance, etc.) à des données intervalles.

Mots-clés :

analyse factorielle, analyse en composantes principales, analyse des correspondances multiples, intervalle, imprécision, variation, contraintes de domaines, codage flou, analyse des données symboliques.

ABSTRACT

Factorial analysis aims to extract and to visualize the main trends of a data set. The aim of this work is to extend principal component and correspondence analysis to interval data. In the first part of this work we present an extension of principal component analysis to interval data, with or without field constraints. We propose two novel methods called the vertices method and the centers method corresponding, respectively, to the within/between analysis and the between analysis. These methods are first compared in a probabilistic case, then a link is made between the vertices method, the Statis method and the discriminent factorial analysis. In the second part of this work, we extend the multiple correspondence method to interval data. To do so, we propose three new fuzzy coding techniques for interval variables: cross-coding, vertices coding and coding without decomposition. The first two are based on the decomposition of interval variables into numerical ones. The third technique is based on the extension of classical coding tools (histogram, probabilistic distribution function, membership function, etc.) to interval distribution.

key-words:

factorial analysis, principal component, multiple correpondence, data analysis, interval data, imprecise data, fuzzy coding, symbolic data analysis.

Remerciements

Ce travail doit beaucoup à la qualité du cadre de recherche dont j'ai bénéficié, aussi bien à l'Inria de Rocquencourt qu'au laboratoire du Lise-Ceremade de l'université Paris-IX Dauphine, ainsi qu'aux nombreux chercheurs qu'il m'a été permis de rencontrer.

Je remercie très vivement mes directeurs de thèse le Professeur Pierre Cazes et le Professeur Edwin Diday.

Je remercie, Monsieur Edwin Diday de m'avoir proposé d'effectuer une thèse sous sa direction, de ses remarques judicieuses, de l'encouragement constant et de la confiance qu'il m'a temoigné durant ma thèse.

Je ne saurais également jamais remercier assez, Monsieur Pierre Cazes pour ses critiques et remarques pertinentes, pour les nombreuses lectures plus que minutiueuses de ce travail de thèse et pour ses encouragements et son soutien permanent.

Je remercie particulièrement, le Professeur Yves Schektman, pour la collaboration très enrichissante et pour tout l'intérêt et l'enthousiasme qu'il a porté à ce travail.

Je remercie vivement, le Professeur Ludovic Lebart, pour l'honneur qu'il m'a fait de présider ce jury de thèse et pour les conseils qui ont permis l'amélioration de ce manuscrit.

Je remercie vivement, le Professeur Gilbert Saporta, pour l'honneur qu'il m'a fait d'accepter d'être rapporteur de mon travail de thèse, pour sa disponibilité et pour les nombreux conseils et critiques qu'il m'a prodigués.

Je remercie vivement, le Professeur Gérard Govaert, pour l'honneur qu'il m'a fait de rapporter ma thèse. Je lui suis particulièrement reconnaissante, pour les nombreuses critiques et améliorations suggérées.

Merci également à Yves Lechevalier, Directeur de recherche à l'Inria, de m'avoir si bien accueilli ces longues années et pour les riches discussions que nous avons eu que ce soit en statistique ou en informatique.

Je ne peux oublier de remercier Myriame Touati et Mireille Summa pour tous les efforts dont elles ont fait preuve pour la gestion du laboratoire Lise-Ceremade avec professionnalisme. Je remercie également Christian Desarménien (Inria Rocquencourt) pour les nombreuses fois où j'ai fait appel à ses compétences en informatique.

Je tiens également à remercier tous mes collègues et amis. Je pense en particulier à : Véronique, Emmanuel, Fred, Ghazi, Barnabé, Younès, Géraldine, Mounir, Halima, Huyen, Moufida, Edgar, Marc, Souleymane, Chafiq, Brieuc et tous les autres avec qui j'ai eu des discussions enrichissantes et agréables. J'ai eu un grand plaisir à travailler ou à discuter avec divers chercheurs étrangers que j'aimerais citer ici : P. Nagabhushan, R. Verde, D. Defays, L. Biard, C. Lauro et A. Gordon.

Je tiens a remercier mon époux pour sa patience, son aide et ses conseils inestimables. Mille mercis à Valérie pour la grande patience dont elle a fait preuve à la relecture de mon manuscrit.

J'aimerais enfin remercier infiniment et affectueusement mes parents.

À Halim

Table des matières

Introduction **1**

1 L'ACP sur des données intervalles **9**
1.1 Introduction 11
1.1.1 L'analyse en composantes principales 11
1.1.2 Données du problème 11
1.1.3 Objectifs 12
1.2 La méthode des sommets 12
1.2.1 Les données 12
1.2.2 Le nuage des hyper-rectangles $N(H)$ 14
1.2.3 Pondération des objets et sémantiques sous-jacentes . . . 15
1.2.4 Pondération des sommets S_k^i 17
1.2.5 Matrice de variance-covariance 20
1.2.6 Description factorielle du nuage d'objets N(H) 21
1.2.7 La projection d'un point de l'hyper-rectangle H_i 22
1.2.8 Le nuage des variables de type intervalle 24
1.2.9 Représentation graphique des objets H_i 25
1.2.10 Généralisation des paramètres d'aide à l'interprétation . 26
1.2.11 Une visualisation qui aide à l'interprétation 28
1.2.12 Algorithme de la procédure itérative de visualisation . . 30
1.2.13 Algorithme de la méthode des sommets 31
1.2.14 L'ACP, un cas particulier de la méthode des sommets . . 32
1.3 Cas de données intervalles dotés de contraintes de domaines . . 33
1.3.1 Contraintes de domaines associées aux objets 33
1.3.2 Pondération sous-contraintes des objets H_i 36
1.3.3 Pondération sous-contraintes des sommets S_k^i 37
1.4 La méthode des centres 38
1.4.1 Les données 38
1.4.2 Le nuage d'objets N(H) 39

1.4.3 Pondération des objets et sémantiques sous-jacentes . . . 40
1.4.4 Matrice de variance-covariance . . . 40
1.4.5 Description factorielle du nuage d'objets . . . 41
1.4.6 Visualisation et interprétation du nuage N(H) . . . 42
1.4.7 Algorithme de la méthode des centres . . . 43
1.4.8 L'Acp, un cas particulier de la méthode des centres . . . 44
1.5 Comparaison de la méthode des sommets et de la méthode des centres . . . 44
1.5.1 Une analyse inter-objets et intra-objets . . . 47
1.5.2 La complexité des calculs . . . 48
1.6 La méthode des sommets du point de vue de la méthode Statis 48
1.7 La méthode des sommets du point de vue de l'analyse factorielle discriminante . . . 50
1.8 Approche probabiliste . . . 50
1.8.1 D'autres distributions à l'intérieur des intervalles . . . 50
1.8.2 La méthode des sommets et des centres du point de vue probabiliste . . . 54
1.8.3 Caractéristiques de tendance centrale et de dispersion des variables X_j . . . 55
1.8.4 Comparaison de la méthode des sommets et de la méthode des centres dans une approche probabiliste . . . 56
1.9 Approche concurrente . . . 60
1.9.1 La réduction de dimension sur des données intervalles . . 61
1.9.2 Application de l'algorithme dans le cas d'une Acp . . . 67
1.9.3 Discussion . . . 70
1.10 Une méthode symbolique de génération de classes d'hyper-rectangles 71

2 Codage flou des variables intervalles en vue d'une Acm 75
2.1 Introduction . . . 77
2.1.1 Le codage des données en Acm . . . 77
2.1.2 L'Acm sous codage disjonctif . . . 81
2.1.3 Le codage flou . . . 83
2.1.4 L'Acm sous codage flou . . . 84
2.1.5 Positionnement du travail . . . 86
2.2 Codage croisé d'une variable de type intervalle . . . 88
2.2.1 Décomposition de la variable intervalle . . . 88
2.2.2 Découpage d'une variable intervalle . . . 89
2.2.3 Fonctions d'appartenance associées aux classes d'intervalles . . . 91

2.2.4 Quelques cas particuliers de codage croisé 94
2.2.5 Exemple 97
2.3 Codage par sommets d'une variable de type intervalle 101
2.3.1 Décomposition des objets 102
2.3.2 Codage des sommets 103
2.3.3 Exemple 103
2.4 Codage des variables intervalles sans décomposition 105
2.4.1 Recouvrement, effectif et densité d'une classe d'intervalles 105
2.4.2 Histogramme d'une distribution d'intervalles 110
2.4.3 Fonction de répartition d'une distribution d'intervalles . 111
2.4.4 Généralisation des techniques de découpage aux variables intervalles 115
2.4.5 Mesures de dissimilarité ou de distance entre intervalles . 116
2.4.6 Propriétés des fonctions d'appartenance associées aux classes d'intervalles 126
2.4.7 Fonction d'appartenance fondée sur la distance de Moore 137
2.4.8 Comparaison des trois techniques de codage d'une variable de type intervalle 137

3 Applications 139
3.1 Application en reconnaissances de visages 140
3.1.1 Présentation de l'application 141
3.1.2 Description des données 141
3.1.3 Résultats dans le cadre de la méthode des sommets . . . 142
3.1.4 Résultats dans le cadre de la méthode des centres 150
3.2 Application en statistique officielle sur la confidentialité des données 155
3.2.1 Introduction 155
3.2.2 Masquage des données par *micro-agrégation* 155
3.2.3 Micro-agrégation par intervalle 156
3.2.4 Comparaisons des variances 157
3.2.5 Application à des statistiques européennes sur le chômage 158
3.3 Application sur les données *Iris* 163
3.3.1 Construction des données intervalles 163
3.3.2 Résultats dans le cadre de la méthode des sommets . . . 164
3.4 Application du codage croisé en Acm 168
3.4.1 Description des données 168
3.4.2 Codage linéaire croisé 168
3.4.3 Résultats et Interprétation 171

3.5 Application du codage par sommets en Acm 175
3.5.1 Codage par sommets 175
3.5.2 Résultats et interprétation 177
3.6 Application du codage sans décomposition en Acm 181
3.6.1 Découpage des variables intervalles par histogramme . . 181
3.6.2 Codage basé sur la distance de Moore 182
3.6.3 Résultats et interprétation 183

Conclusion **187**

Introduction

Les méthodes d'analyse de données étudient des objets dont les descriptions peuvent être représentées par un tableau, où chaque objet (ligne du tableau) est décrit par une valeur unique pour chaque variable (colonne du tableau).

Les descriptions soumises aux méthodes d'analyse de données, qu'elles soient fournies par un expert, extraites de bases de données ou d'une base de connaissance, sont de plus en plus complexes. Par données complexes on entend : des intervalles (point de fusion $= [15^o, 17^o]$), des ensembles de valeurs (nombre de dents $\in \{2, 3\}$), des contraintes de domaines (point d'éclair $> 110^oC$), des dépendances entre descripteurs (longueur des pétales ayant 3 à 4 fois la longueur des sépales), les valeurs peuvent être entachées d'incertitude ou de doute, des liens de composition ou de hiérarchie peuvent être exprimés entre des descriptions.

L'approche adoptée jusqu'à présent pour analyser de telles descriptions consiste à tronquer les données pour qu'elles puissent entrer dans un schéma tabulaire. Ces transformations fournissent, en général, des descriptions de tendances centrales : un intervalle est remplacé par son centre, un ensemble de valeurs est codé par la moyenne ou le mode, un ensemble de valeurs entachées d'incertitude ou de doute est remplacé par la valeur la plus probable, etc. Quant aux dépendances et aux liens de compositions ou de hiérarchies entre les descripteurs ou entre les objets, ils sont souvent ignorés.

Analyser et étudier des descriptions plus fiables, étendre l'analyse de données à des données plus riches, mettre les méthodes d'analyse de données à la disposition des données et non l'inverse, telle est l'essence même de **l'analyse de données symboliques**.

L'analyse de données symboliques

L'analyse de données symboliques fut proposée par E. DIDAY [Diday89b], [Diday91], [Diday95]. Le but de l'analyse de données symboliques est d'étendre la problématique, les méthodes et algorithmes d'analyse de données à des descriptions plus complexes dites symboliques.

Une **description symbolique** se distingue de celle classiquement traitée en analyse de données par :

1. Chaque variable peut prendre des **valeurs multiples** pour un même objet.
   Exemple : couleur $\in \{$*jaune, marron,* $\{$*noir, marron*$\}\}$ *pour exprimer que la couleur d'un cèpe peut être jaune ou marron ou noir et marron ; point de fusion* $\in [15^o, 17^o]$ *pour exprimer que la température de fusion varie entre 15 et 17 degrés.*

2. L'expression de liens ou de **contraintes de domaines** entre les variables et/ou les objets.
   Exemple : la contrainte de domaine entre la taille et la couleur des cèpes est : **si** *couleur=jaune,* **alors** *la taille varie entre [0,7] et* **si** *couleur=marron,* **alors** *la taille varie entre [7, 15].*

3. Une description symbolique est une description en compréhension d'une classe de descriptions qui en constituent **l'extension**.
   Exemple : la description définie par couleur $\in \{$*blanc, jaune*$\}$ *a pour* **extension** *toutes les descriptions définies soit par couleur* $\in \{$*blanc*$\}$ *soit par couleur* $\in \{$*jaune*$\}$.

4. L'ensemble des descriptions symboliques est muni d'opérateurs de **généralisation** et de **spécialisation** tenant compte des sémantiques des domaines auxquels ils s'appliquent. Le processus de généralisation et de spécialisation permet d'agréger un ensemble de descriptions symboliques en une description symbolique *généralisante*, respectivement, de synthétiser un ensemble de descriptions symboliques en une *spécifique*.

Quelques travaux réalisés en analyse de données symboliques

Nous ne décrivons ici que très brièvement quelques travaux récents réalisés en analyse de données symboliques.

En classification automatique, les premiers travaux réalisés par P. Brito [Brito91] concernant la construction de pyramides symboliques ont ensuite été repris par E. Mfoumoune [Mfoumoune98] avec la prise en compte en entrée de descriptions probabilistes. Toujours en classification automatique de descriptions symboliques, signalons les travaux de M. Chavent [Chavent97] sur le thème des méthodes divisives et ceux de G. Polaillon [Polaillon et al.96] sur la construction et la réduction de treillis de Gallois.

Les travaux liés aux méthodes explicatives en analyse des données symboliques ont porté essentiellement sur les techniques de discrimination par arbre ou segmentation. Ainsi, outre l'approche par graphe d'identification réalisée par J. Lebbe et R. Vignes [Lebbe et al.91], on peut citer les travaux de E. Perinel [Perinel96] sur l'extension de l'algorithme de segmentation classique à des données empreintes d'imprécision.

Dans le domaine de la discrimination se pose souvent le problème de la sélection de variables. Ces techniques de sélection de variables sont particulièrement importantes dans la mesure où elles permettent très souvent de réduire sensiblement la complexité d'un problème décisionnel. Dans ce domaine, D. Ziani [Ziani96] a proposé une extension de la méthode minset ([Vignes91]) à des objets symboliques probabilistes.

Dans une optique plus descriptive ou exploratoire, citons enfin les travaux de Y. Hillali [Diday et al.96], [Hillali98] relatifs à la construction d'histogrammes de capacité à partir d'objets probabilistes ; ceux de F. Decarvalho [Carvalho92], [Carvalho95] sur les méthodes descriptives en analyse des données symboliques ; ou encore ceux de V. Stephan [Stephan98], dont l'objectif est de résumer les résultats d'une requête Sql, soumise à une base de données relationnelle, par un ensemble d'objets symboliques en utilisant des procédés de *généralisation/spécialisation.*

Positionnement du problème

L'analyse factorielle tient une place primordiale parmi les méthodes d'analyse de données. Elle est largement utilisée dans de nombreux domaines dont en analyse d'images et en reconnaissance de formes. Elle permet d'extraire, à partir de données nombreuses, les tendances les plus marquantes. À l'aide de représentations graphiques, elle visualise des groupements, des oppositions, des tendances impossibles à discerner directement sur un tableau de données. La nature des données traitées, les poids, la métrique, introduisent des variantes au sein des méthodes d'analyse factorielle. On dinstingue principalement : l'analyse en composantes principales, l'analyse des correspondances simples et multiples.

L'analyse en composantes principales s'applique à des tableaux de mesures dont les colonnes représentent les variables mesurées (quantitatives) et les lignes les individus sur lesquels ont été effectués ces mesures. Dans le cas de tableaux portant sur des variables hétérogènes (qualitatives et/ou quantitatives) l'analyse des correspondances multiples est utilisée à la suite de codages des variables initiales en des variables binaires ou floues.

Dans la démarche "symbolique/numérique", qui tend vers le traitement de données plus riches, une extension naturelle est celle que constituent les intervalles pour les données numériques. Les intervalles permettent de décrire de manière plus fiable l'information d'imprécision, de doute ou de variation sur un domaine.

Origines des intervalles

Souvent les données numériques manipulées en statistique ou en analyse de données cachent des données intervalles. En effet, toute donnée mesurée est un intervalle de la forme $[x \pm \delta]$ où, x est le résultat de la mesure et δ l'imprécision dûe à l'instrument de mesure.

D'autre part, l'avancé des systèmes informatiques prévoit actuellement de nouveaux types (au sens informatique) dont les intervalles, permettant ainsi au bases de données et à l'intelligence artificielle de décrire et de gérer des applications incluant des intervalles.

Les intervalles peuvent également être obtenus par construction selon une théorie rattachée à un domaine, et exprimer une sémantique particulière.
Par exemple, un intervalle de confiance, construit en statistique, exprime le domaine de variation d'un paramètre avec un degré de confiance donné. En théorie des possibilités, connaissant les degrés de croyance $a = d(A)$ et $b = d(B)$ d'un expert relativement aux faits A et B, alors le degré de croyance $d(A \wedge B)$ du fait $A \wedge B$ est l'intervalle $[min(a+b-1, 0), min(a, b)]$. En classification, un intervalle décrivant une classe est construit à partir d'un ensemble de valeurs observées pour cette classe (nombre d'étamines du genre "Jaquini" varie entre 16 et 20) etc.

De par l'intérêt que revêt la notion d'intervalle, nous nous sommes intéressés dans notre travail de thèse à l'extension de l'analyse en composantes principales et de l'analyse des correspondances multiples à des tableaux portant sur des données intervalles. Dans cette même problématique citons deux approches.

Il y a d'abord, les modèles à effets fixes proposés par Caussinus [Caussinus92] permettant l'application de l'analyse en composantes principales à des données entachées d'erreurs. Dans ces modèles, on suppose que les lignes (individus) du tableau sont des vecteurs aléatoires indépendants. Chaque ligne est décomposable en la somme de deux termes : un effet fixe (la moyenne) et un effet aléatoire (l'erreur autour de la moyenne). On fait l'hypothèse que les moyennes appartiennent à un sous-espace inconnu de dimension fixé.

D'autre part, Nagabushan [Nagabhushan97] propose une extension de l'analyse en composante principale à des données intervalles basée sur les dérivées partielles. Cette méthode est présentée en détail dans un paragraphe ultérieur.

Un mouvement global vers le traitement de la connaissance

Les évolutions réalisées en informatique : saisie automatique de données, grandes capacités mémoires, de puissants processeurs etc. et la propension à automatiser de plus en plus de tâches font apparaître un nombre grandissant d'applications nécessitant des descriptions plus complexes.

Face à la difficulté de modéliser et de gérer de telles données de nombreux travaux ont été entrepris dans d'autres domaines de traitement de l'information tels que les bases de données et l'intelligence artificielle.

Les bases de données et la connaissance

En base de données, face à l'échec du schéma relationnel à modéliser et à traiter des applications complexes, plusieurs solutions sont proposées.

Il y a eu d'abord le modèle *relationnel étendu* qui a permis d'ajouter de nouveaux types de données : des textes longs, des intervalles, des ensembles de valeurs, etc. Il y eut ensuite l'apparition des *modèles sémantiques* qui ont permis essentiellement la description de liens entre objets : associations, agrégations, généralisations, spécialisations. Finalement a émergé dans la première moitié des années quatre-vingts un courant alliant les langages dits "orientés objets", possédant un pouvoir de modélisation plus riche que les langages de programmation traditionnels, et un modèle de données comparable aux modèles sémantiques. Ces systèmes constituent actuellement une réponse à de nombreux problèmes complexes ; ils permettent, par exemple, la saisie et la manipulation de données de type intervalle, de type ensemble, de définir des dépendances logiques ou des liens hiérarchiques entre des objets. Il apparaît ainsi un décalage entre le type de données accumulées dans de telles bases et les méthodes traditionnelles de la statistique mieux adaptées à traiter des données standards.

L'intelligence artificielle et les bases de connaissance

En intelligence artificielle, la représentation de la connaissance a suscité de nombreux travaux. Il y a eu tout d'abord les représentations fondées sur la logique des prédicats et sur la programmation logique (langage Prolog) ; viennent ensuite les représentations procédurales qui permettent de bien représenter les inter-relations entre fragments de connaissances. Finalement, la représentation mixte qui combine les deux modes de représentation précédents et qui est fondée sur la notion d'objets structurés (frames, scripts, schémas, etc.). Un objet structuré peut être décrit par des données intervalles, des ensembles, prendre des valeurs par défaut et être rattaché à des procédures.

Un nouveau domaine d'extraction des connaissances : "Data-Mining"

Les évolutions réalisées en informatique ont comme conséquence l'augmentation massive des données dans les bases. Face à ce déluge de données, créant un grand décalage entre la génération des données et leurs analyse, un problème fondamental se pose : comment en extraire des informations utiles dans un but explicatif ou décisionnel. Par informations utiles, on entend de nouveaux concepts, de nouvelles règles, non existant physiquement dans la base, ni déduits à partir de connaissances a priori. C'est ce qui constitue l'objectif du nouveau domaine de recherche "Data-Mining" dit encore "Knowledge Data Discovery" (KDD). Le domaine de "Data-Mining" est situé au confluent des bases de données (orientées objets), de l'apprentissage symbolique automatique (branche de l'intelligence artificielle) et de l'analyse de données.

Pour extraire des informations non connues a priori et ne pouvant pas être obtenues par un processus simple, à partir des données de la base, le domaine du "Data-Mining" utilise les fonctions de stockage et de gestion (modification, suppression, etc.) des données fournies par des bases de données, les techniques d'induction / déduction fournies par les méthodes d'apprentissage symbolique automatique, et les méthodes de statistique et d'analyse de données.

Dans ce nouveau domaine de recherche en pleine expansion, alors que les bases de données et l'intelligence artificielles gèrent des données plus riches et plus complexes, les méthodes d'analyse de données restent limitées à des données numériques ou qualitatives.

Présentation du travail

Dans le premier chapitre, on présente **une extension de l'analyse en composantes principales** à des données de type **intervalle**, munies éventuellement de **contraintes**. On propose deux nouvelles approches : *la méthode des sommets* et *la méthode des centres.*

On détaille chacune de ces méthodes : recodage, pondération, inertie prise en compte, description factorielle, visualisation, etc. Les algorithmes des méthodes proposées sont fournis en pseudo-code. On présente, dans le cas de la méthode des sommets, la façon dont des contraintes de domaines définies sur les données intervalles peuvent être prises en compte et visualisées dans les

plans factoriels. Toujours dans le cadre de la méthode des sommets, on propose une procédure itérative qui fournit la visualisation des objets à différents niveaux de qualité de représentation.

D'une part, on confronte et on compare, dans un cadre probabiliste, la méthode des sommets et la méthode des centres. D'autre part, on établit un rapprochement entre la méthode des sommets, la méthode STATIS et l'analyse factorielle discriminante. Finalement, on présente une méthode concurrente que l'on discute.

Dans le second chapitre, on s'intéresse à **l'analyse des correspondances multiples**. On propose trois nouvelles techniques de codage des variables de type intervalle : *le codage croisé*, *le codage par sommets* et *le codage sans décomposition*. On présente en détail le principe de chaque codage, ses propriétés ainsi que les interprétations sous-jacentes en analyse des correspondances multiples.

Dans le dernier chapitre, on met en application les techniques et méthodes proposées. La première application concerne un problème de reconnaissance de visage. La deuxième application porte sur le problème de la confidentialité des données. On montre, à travers un jeu de données sur le chômage en Europe, l'intérêt des intervalles et de la méthode des sommets pour le codage de données confidentielles. Pour vérifier la cohérence des résultats fournis par les méthodes proposées, on utilise dans la troisième application des données connues : *les Iris de* FISHER. Les techniques de codage en analyse des correspondances multiples sont également testées et appliquées à un jeu de données connu.

Les algorithmes mentionés dans cette thèse ont été totalement implémentés en C++, sous un environnement UNIX ; ils ne font appel à aucun outil ou logiciel extérieur. Certains de ces modules sont également fonctionnels sous WINDOWS'95.

Chapitre 1

L'ACP sur des données intervalles

Résumé du chapitre

L'analyse en composantes principales est l'une des méthodes d'analyse de données les plus largement utilisées. On propose d'étendre cette méthode à un type de donnée complexe, fréquemment rencontré dans les descriptions : les intervalles.

On propose deux approches : la méthode des sommets et la méthode des centres. L'inertie prise en compte par chacune de ces méthodes est analysée puis comparée. Les paramètres d'aide à l'interprétation sont généralisés aux données intervalles. La visualisation ponctuelle des individus, dans les plans factoriels, se traduit ici par une visualisation de rectangles, segments ou points.

On propose une procédure itérative qui fournit une visualisation des objets à différents niveaux de qualité de représentation. On montre comment la méthode des sommets, munie d'un système de pondération adéquat, permet de tenir compte de contraintes de domaines définies sur les données intervalles et de les visualiser dans les plans factoriels.

D'une part, on confronte dans un cadre probabiliste la méthode des sommets et la méthode des centres. D'autre part, on établit un rapprochement entre la méthode des sommets, la méthode STATIS *et l'analyse factorielle discriminante.*

On discute, finalement, une méthode concurrente de réduction de dimension fondée sur la dérivation partielle.

1.1 Introduction

1.1.1 L'analyse en composantes principales

Étant donné un ensemble d'objets décrits par des variables quantitatives, les problèmes souvent confrontés en analyse de données sont :

- *La réduction de dimension* : on cherche à décrire l'ensemble des objets par un nombre réduit de variables.

- *L'exploration de la structure principale des objets* : on cherche à extraire la tendance principale de la distribution des objets : les principaux groupements d'objets, les objets qui s'opposent ou qui sont similaires, etc.

Pour répondre à ces deux problèmes, on a souvent recours à l'analyse en composantes principales (ACP). L'ACP est une méthode descriptive qui permet de fournir une représentation approchée du nuage d'objets dans un nouvel espace de dimension plus faible. D'une part, les objets sont décrits par un nombre plus faible de nouvelles variables dites composantes principales ; d'autre part, la structure principale des objets est présentée à travers plusieurs plans factoriels, où les proximités entre les objets sont appréciées à l'aide de paramètres dits d'aide à l'interprétation.

1.1.2 Données du problème

Soit $H_1, \ldots, H_m$ m objets décrits par q variables $X_1, \ldots, X_q$ de type intervalle :

$$X_H = \begin{pmatrix} H_1 \\ \vdots \\ H_m \end{pmatrix} = \begin{pmatrix} [\underline{x_{11}}, \overline{x_{11}}] & \cdots & [\underline{x_{1q}}, \overline{x_{1q}}] \\ \vdots & \ddots & \vdots \\ [\underline{x_{m1}}, \overline{x_{m1}}] & \cdots & [\underline{x_{mq}}, \overline{x_{mq}}] \end{pmatrix} \tag{1.1}$$

où $[\underline{x_{ij}}, \overline{x_{ij}}]$ est l'intervalle pris par la variable X_j pour l'objet H_i et $\underline{x_{ij}}$, $\overline{x_{ij}}$ sont, respectivement, la plus petite et la plus grande valeur prises par la variable X_j pour l'objet H_i. Un intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$ est dit trivial s'il est réduit à

une valeur : $x_{ij} = \overline{x_{ij}}$. Une variable X_j est dite de type intervalle s'il existe au moins un objet H_i $(i = 1..m)$ tel que $[\underline{x_{ij}}, \overline{x_{ij}}]$ soit un intervalle non trivial.

La matrice X_H constitue une généralisation des tableaux classiques (individus x variables). En effet, dans le cas particulier où tous les intervalles $[\underline{x_{ij}}, \overline{x_{ij}}]$ sont triviaux, la matrice X_H donne la description des m objets par q variables numériques.

1.1.3 Objectifs

Notre objectif est de généraliser l'ACP à un ensemble d'objets décrits par des variables de type intervalle. Cette généralisation doit assurer, d'une part, les fonctions principales d'une ACP classique : réduction de dimension et extraction de la structure principale des objets. D'autre part, elle doit restituer l'information de variation ou d'imprécision introduite par ces variables. Finalement, dans le cas particulier d'un tableau classique (individus x variables) les résultats fournis par l'ACP généralisée doivent coïncider avec ceux issus d'une ACP classique.

1.2 La méthode des sommets

1.2.1 Les données

Soit q_i le nombre d'intervalles non triviaux dans la description d'un objet H_i. Un tel objet peut être visualisé dans l'espace de description défini par les variables $X_1, \ldots, X_q$, par un hyperparallélépipède-rectangle (qu'on appellera plus simplement hyper-rectangle) à $n_i = 2^{q_i}$ sommets.

La longueur des côtés de l'hyper-rectangle est donnée par l'étendue des intervalles associés à chaque variable de description. On note n le nombre total de sommets des hyper-rectangles tout objet confondu :

$$n = \sum_{i=1}^{m} 2^{q_i} = \sum_{i=1}^{m} n_i$$

La figure 1.1 suivante représente un nuage d'objets (hyper-rectangles), décrits par trois variables de type intervalle.

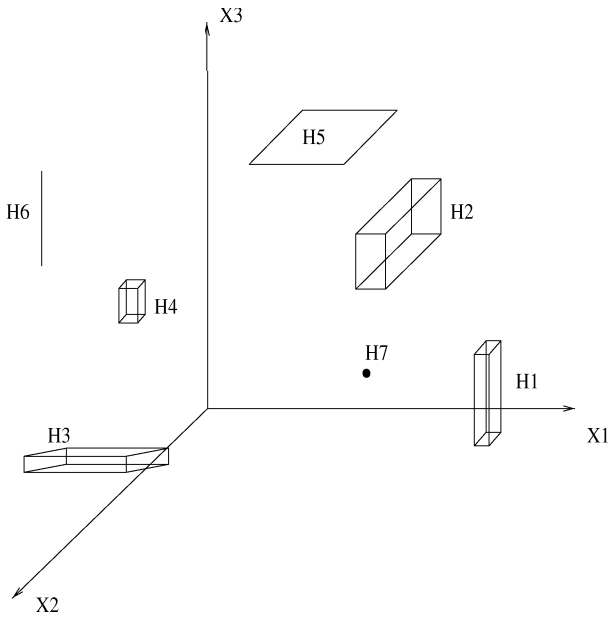


FIG. 1.1: *Nuage d'objets décrits par trois variables de type intervalle*

Dans ce qui suit, un hyper-rectangle peut aussi bien désigner un point ($q_i = 0$, cas de H7), un segment ($q_i = 1$, cas de H6) ou un rectangle ($q_i = 2$, cas de H5). On propose de décrire un objet H_i défini en 1.1 par la matrice suivante :

$$X_{H_i} = \begin{pmatrix} x_{S_1^i 1} & \cdots & x_{S_1^i q} \\ \vdots & \ddots & \vdots \\ x_{S_{n_i}^i 1} & \cdots & x_{S_{n_i}^i q} \end{pmatrix} \tag{1.2}$$

où $\{S_1^i, \ldots, S_{n_i}^i\}$ est l'ensemble des sommets de l'hyper-rectangle H_i et $x_{S_k^i j}$ est la valeur prise par la variable X_j pour le sommet k de l'hyper-rectangle H_i.

Soit la description, par exemple, de l'objet H_5 (visualisé dans la figure 1.1) par trois variables de type intervalle :

$$H_5 = \Big([\underline{x_{51}}, \overline{x_{51}}], [\underline{x_{52}}, \overline{x_{52}}], [\underline{x_{53}}, \overline{x_{53}}]\Big)$$

avec $x_{53} = \underline{x_{53}} = \overline{x_{53}}$.

La description de H_5 à travers ses sommets est donnée par la matrice X_{H_5} suivante :

$$X_{H_5} = \begin{pmatrix} \underline{x_{51}} & \underline{x_{52}} & x_{53} \\ \underline{x_{51}} & \overline{x_{52}} & x_{53} \\ \overline{x_{51}} & \underline{x_{52}} & x_{53} \\ \overline{x_{51}} & \overline{x_{52}} & x_{53} \end{pmatrix}$$

On note X la matrice des données obtenue en concaténant les matrices X_{H_i} comme suit :

$$X = \begin{pmatrix} X_{H_1} \\ \vdots \\ X_{H_m} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} x_{S_1^1 1} & \cdots & x_{S_1^1 q} \\ \vdots & \ddots & \vdots \\ x_{S_{n_1}^1 1} & \cdots & x_{S_{n_1}^1 q} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} x_{S_1^m 1} & \cdots & x_{S_1^m q} \\ \vdots & \ddots & \vdots \\ x_{S_{nm}^m 1} & \cdots & x_{S_{nm}^m q} \end{bmatrix} \end{pmatrix} \tag{1.3}$$

1.2.2 Le nuage des hyper-rectangles $N(H)$

À tout objet $H_i \in N(H)$ on fait correspondre n_i vecteurs $\{x_{S_1^i}, \ldots, x_{S_{n_i}^i}\}$ de $\Re^q$ tels que :

$$x_{S_k^i} = \left(x_{S_k^i 1}, \ldots, x_{S_k^i q}\right) \tag{1.4}$$

Chaque objet H_i constitue un sous-nuage de points formé par les sommets de l'hyper-rectangle associé. Le nuage $N(H)$ des objets défini en 1.1 est représenté dans la figure 1.2.

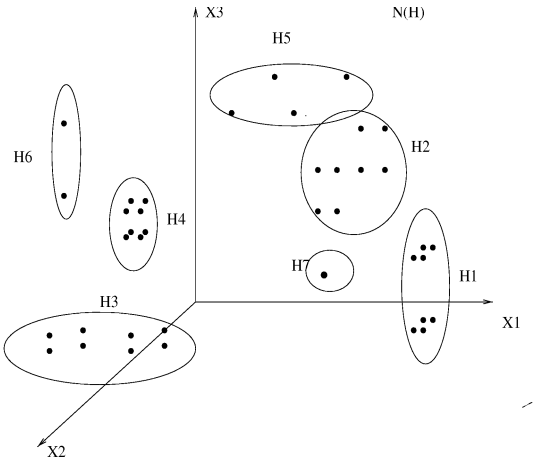


FIG. 1.2: *Nuage N(H) des sommets des hyper-rectangles*

Étant donné un nuage d'hyper-rectangles, l'ACP généralisée aux intervalles, comme l'ACP classique, doit renvoyer une succession de plans factoriels respectant aux mieux les proximités entre les hyper-rectangles.

1.2.3 Pondération des objets et sémantiques sous-jacentes

On adopte dans ce qui suit les notations suivantes :

p_i : le poids de l'objet H_i
$p_{S_k^i}$: le poids du sommet S_k de l'objet H_i
$V(H_i)$: le volume de l'hyper-rectangle H_i calculé comme suit :

$$V(H_i) = \prod_{(\overline{x_{ij}} \neq \underline{x_{ij}})} (\overline{x_{ij}} - \underline{x_{ij}})$$

Les poids p_i des objets et $p_{S_k^i}$ des sommets vérifient les relations suivantes :

$$p_i = \sum_{k=1}^{n_i} p_{S_k^i}$$
$$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \sum_{k=1}^{n_i} p_{S_k^i} = 1$$

Pondération des objets

Le système de pondération le plus fréquent en Acp consiste à attribuer aux objets des poids équivalents, soit :

$$\forall \ i = 1..m \ \ p_i \ = \ \frac{1}{m}$$

On peut envisager d'autres types de pondération tenant compte de la sémantique des intervalles manipulés. Dans le cas des intervalles exprimant de la variation ou de l'imprécision, on distingue deux types de pondération. Le premier système de pondération accorde des poids importants aux objets volumineux (grande variation/faible précision). Le deuxième système accorde des poids importants aux objets non volumineux (faible variation/grande précision).

a) Pondérations proportionnelles aux volumes des hyper-rectangles

On propose, par exemple, d'attribuer à chaque objet H_i un poids p_i proportionnel à l'amplitude de la variation ou de l'imprécision introduite par celui-ci, soit :

$$p_i \ = \ \frac{V(H_i)}{\sum_{i=1}^{m} V(H_i)}$$

Un objet H_i est d'autant plus pesant dans l'analyse (dans la détermination des axes factoriels) que son volume est important. Les objets réduits à un point auront un poids nul.

b) Pondérations inversement proportionnelles aux volumes des hyper-rectangles

À l'opposé, si l'on veut attribuer des poids élevés aux objets représentés par des hyper-rectangles de faible volume (faible variation/grande précision), on adopte la pondération suivante :

$$p_i \ = \ \frac{\left(1 - \frac{V(H_i)}{\sum_{i=1}^{m} V(H_i)}\right)}{\sum_{i=1}^{m} \left(1 - \frac{V(H_i)}{\sum_{i=1}^{m} V(H_i)}\right)}$$

Ce type de pondération est plus approprié à des intervalles exprimant de l'imprécision. En effet, plus le volume d'un objet est faible, plus il est considéré comme précis et plus son poids est élevé dans l'analyse. À l'inverse, un objet de volume important est considéré comme peu précis et aura, par conséquent, un poids faible dans l'analyse.

1.2.4 Pondération des sommets S_k^i

Connaissant le poids p_i de l'objet H_i, le problème de la pondération des sommets S_k^i revient à déterminer la proportion du poids p_i à affecter à chaque sommet. On propose de répartir le poids p_i de chaque objet H_i en tenant compte des distributions empiriques à l'intérieur des intervalles. Celles-ci sont supposées indépendantes.

La valeur de référence x_{ij}^o

On note $x_{ij}^o \in [\underline{x_{ij}}, \overline{x_{ij}}]$ la valeur de référence de l'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$. Rappelons qu'un intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$ peut être le résumé d'un ensemble de valeurs $x_{ij}^1, \ldots, x_{ij}^r$ pris par la variable X_j pour l'objet H_i avec $\underline{x_{ij}} = \min_{k=1..r} x_{ij}^k$ et $\overline{x_{ij}} = \max_{k=1..r} x_{ij}^k$.

Si l'on dispose de l'ensemble des valeurs à partir desquelles l'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$ a été construit, la valeur de référence x_{ij}^o peut désigner : la moyenne des différentes valeurs observées, le mode, une valeur $x_{ij}^k \in [\underline{x_{ij}}, \overline{x_{ij}}]$ choisie par l'expert etc. Dans le cas contraire, on considère le centre de l'intervalle comme valeur de référence.

Les coefficients de pondération des bornes des intervalles

Connaissant la valeur de référence x_{ij}^o de chaque intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$, il s'agit de calculer les coefficients $\underline{p_{ij}}$ et $\overline{p_{ij}}$ des bornes $\underline{x_{ij}}$ et $\overline{x_{ij}}$ de l'intervalle tel que :

$$\underline{p_{ij}} + \overline{p_{ij}} = 1 \tag{1.5}$$

$$\underline{p_{ij}}\,\underline{x_{ij}} + \overline{p_{ij}}\,\overline{x_{ij}} = x_{ij}^o \tag{1.6}$$

Pondération des sommets

Après avoir déterminé les coefficients de toutes les bornes des intervalles décrivant l'objet H_i et sachant les distributions à l'intérieur des intervalles indépendantes, les poids $p_{S_k^i}$ des sommets S_k^i sont obtenus comme suit :

$$\begin{array}{rrcll} & p_{S_k^i} & = & \left(\Pi_{j=1}^{q_i} p(x_{S_k^i j})\right)\ p_i & \\ \text{où} & p(x_{S_k^i j}) & = & \underline{p_{ij}} & \text{si } x_{S_k^i j} = \underline{x_{ij}} \\ & & = & \overline{p_{ij}} & \text{si } x_{S_k^i j} = \overline{x_{ij}} \end{array}$$

On vérifie, aisément, que $\sum_{k=1}^{n_i} p_{S_k^i} = p_i$.

Exemple

Considérons, à titre d'exemple, la pondération des sommets de l'objet H_i de poids p_i décrit dans la figure 1.3 suivante :

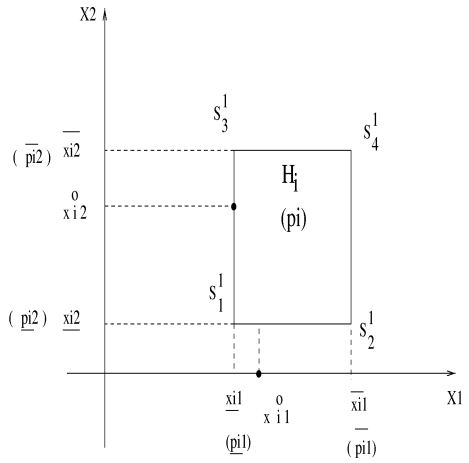


Fig. 1.3:

$\underline{p_{ij}}$ et $\overline{p_{ij}}$ sont les coefficients des bornes $\underline{x_{ij}}$ et $\overline{x_{ij}}$.

Les poids des sommets S_k^i (k=1..4) sont définis comme suit :

$$p_{S_1^i} = \underline{p_{i1}}\ \underline{p_{i2}}\ p_i$$

$$\begin{aligned} p_{S_2^i} &= \overline{p_{i1}}\ \underline{p_{i2}}\ p_i \\ p_{S_3^i} &= \underline{p_{i1}}\ \overline{p_{i2}}\ p_i \\ p_{S_4^i} &= \overline{p_{i1}}\ \overline{p_{i2}}\ p_i \end{aligned}$$

On vérifie aisément que la somme des poids des sommets est bien égale au poids p_i de l'objet H_i.

Équipondération des sommets

Étudions la pondération des sommets dans le cas où l'on ne dispose d'aucune connaissance sur la distribution à l'intérieur des intervalles. Dans ce cas-là, tous les points de référence sont estimés par les centres des intervalles.

$$\forall\ j = 1..q_i \qquad x_{ij}^o = \frac{\underline{x_{ij}} + \overline{x_{ij}}}{2}$$

soit :

$$\forall\ i = 1..m\ \ j = 1..q_i \qquad \underline{p_{ij}} = \overline{p_{ij}} = \frac{1}{2}$$

Les poids des sommets de l'objet H_i sont alors :

$$\forall\ k = 1..n_i \qquad p_{S_k^i} = \Pi_{j=1}^{q_i} \frac{1}{2}\, p_i = \frac{1}{2^{q_i}}\, p_i = \frac{p_i}{n_i}$$

Dans le cas où tous les points de référence sont situés aux centres des intervalles, le poids p_i de l'objet H_i est réparti équitablement entre tous les sommets. On vérifie alors que :

$$\sum_{i=1}^{n_i} p_{S_k^i} = \sum_{i=1}^{n_i} \frac{p_i}{n_i} = n_i \frac{p_i}{n_i} = p_i$$

La matrice des poids D

On note D la matrice diagonale des poids des sommets, tous objets confondus, à n lignes (n nombre total des sommets) et q colonnes, définie comme suit :

$$D = \begin{pmatrix} p_{S_1^1} & 0 & \dots & & 0 \\ 0 & \ddots & & & \\ & & p_{S_{n_1}^1} & & \\ \vdots & & \ddots & & \vdots \\ & & & p_{S_1^m} & \\ \vdots & & & \ddots & 0 \\ 0 & & \dots & 0 & p_{S_{n_m}^m} \end{pmatrix} \tag{1.7}$$

1.2.5 Matrice de variance-covariance

Supposons, ce qui ne restreint pas la généralité, que les variables sont centrées. La matrice de variance-covariance V^s est définie comme suit :

$$V^s = X^T D X \tag{1.8}$$

La moyenne $\overline{X_j^s}$ de X_j s'écrira donc :

$$\overline{X_j^s} = \sum_{i=1}^{m} \sum_{k=1}^{n_i} p_{S_k^i} x_{S_k^i j} \tag{1.9}$$

$$= \sum_{i=1}^{m} (\underline{\alpha_{ij}}\ \underline{x_{ij}} + \overline{\alpha_{ij}}\ \overline{x_{ij}}) \tag{1.10}$$

où, $\underline{\alpha_{ij}}$ est la somme des poids des sommets de l'objet H_i prenant la valeur $\underline{x_{ij}}$, et $\overline{\alpha_{ij}}$ est la somme des poids des sommets de H_i prenant la valeur $\overline{x_{ij}}$, soit :

$$\underline{\alpha_{ij}} = \sum_{(k=1\ /\ x_{S_k^i j} = \underline{x_{ij}})}^{n_i} p_{S_k^i} = \underline{p_{ij}}\, p_i \tag{1.11}$$

$$\overline{\alpha_{ij}} = \sum_{(k=1\ /\ x_{S_k^i j} = \overline{x_{ij}})}^{n_i} p_{S_k^i} = \overline{p_{ij}}\, p_i \tag{1.12}$$

$$\underline{\alpha_{ij}} + \overline{\alpha_{ij}} = \underline{p_{ij}}\, p_i + \overline{p_{ij}}\, p_i = p_i \tag{1.13}$$

La description de l'objet H_i par la variable X_j fait intervenir deux termes $\underline{x_{ij}}$ et $\overline{x_{ij}}$ chacun 2^{q_i-1} fois. Ayant supposé le tableau X centré, on a $\overline{X_j^s} = 0$. La variance v_{jj}^s de la variable X_j s'exprime comme suit :

$$
\begin{aligned}
v_{jj}^s &= \sum_{i=1}^{m} \sum_{k=1}^{n_i} p_{S_k^i} (x_{S_k^i j})^2 && (1.14)\\
&= \sum_{i=1}^{m} \Big(\sum_{(k=1\ /\ x_{S_k^i j} = \underline{x_{ij}})}^{n_i} p_{S_k^i} \Big) \underline{x_{ij}}^2 + \Big(\sum_{(k=1\ /\ x_{S_k^i j} = \overline{x_{ij}})}^{n_i} p_{S_k^i} \Big) \overline{x_{ij}}^2 && (1.15)\\
&= \sum_{i=1}^{m} (\underline{\alpha_{ij}}\ \underline{x_{ij}}^2 + \overline{\alpha_{ij}}\ \overline{x_{ij}}^2) && (1.16)
\end{aligned}
$$

La covariance v_{jl}^s entre les variables X_j et X_l $(j \neq l)$ est définie comme suit :

$$
v_{jl}^s = \sum_{i=1}^{m} \sum_{k=1}^{n_i} p_{S_k^i} x_{S_k^i j}\ x_{S_k^i l}\ . \qquad (1.17)
$$

La description de l'objet H_i par les variables X_j et X_l $(j \neq l)$ fait intervenir quatre termes $\underline{x_{ij}}\underline{x_{il}}$, $\underline{x_{ij}}\overline{x_{il}}$, $\overline{x_{ij}}\underline{x_{il}}$, $\overline{x_{ij}x_{il}}$ chacun 2^{q_i-2} fois ; sachant que la somme des poids des sommets S_k^i $(k = 1..n_i)$ de coordonnées, par exemple, $x_{S_k^i j} = \underline{x_{ij}}$ et $x_{S_k^i l} = \overline{x_{il}}$ est égale à $p_i\, \underline{p_{ij}}\, \overline{p_{il}}$, alors l'expression de la covariance devient :

$$
\begin{aligned}
v_{jl}^s &= \sum_{i=1}^{m} p_i \left(\underline{p_{ij}}\, \underline{p_{il}}\, \underline{x_{ij}}\, \underline{x_{il}} + \underline{p_{ij}}\, \overline{p_{il}}\, \underline{x_{ij}}\, \overline{x_{il}} + \overline{p_{ij}}\, \underline{p_{il}}\, \overline{x_{ij}}\, \underline{x_{il}} + \overline{p_{ij}}\, \overline{p_{il}}\, \overline{x_{ij}}\, \overline{x_{il}} \right)\\
&= \sum_{i=1}^{m} p_i (\underline{p_{ij}x_{ij}} + \overline{p_{ij}x_{ij}})(\underline{p_{il}x_{il}} + \overline{p_{il}x_{il}})\\
&= \sum_{i=1}^{m} p_i x_{ij}^o x_{il}^o\ .
\end{aligned}
$$

1.2.6 Description factorielle du nuage d'objets N(H)

Pour la détermination des principaux axes d'inertie du nuage d'objets N(H) décrit par la matrice X dans l'espace $\Re^q$ on diagonalise la matrice variance-covariance V^s, la matrice identité étant la métrique adoptée. Soient u_j, λ_j et

PC_j, respectivement, le $j^{ème}$ vecteur propre normé de V^s, la $j^{ème}$ valeur propre associée et la $j^{ème}$ composante principale.

$$PC_j \quad = \quad X.u_j$$

On note F la matrice à n lignes et p colonnes donnant la description factorielle des n sommets des hyper-rectangles dans l'espace défini par les p premières composantes principales $PC_1, \ldots, PC_p$:

$$F \quad = \quad \begin{pmatrix} pc_{S_1^1 1} & \cdots & pc_{S_1^1 p} \\ \vdots & \ddots & \vdots \\ pc_{S_{n_1}^1 1} & \cdots & pc_{S_{n_1}^1 p} \\ \vdots & \ddots & \vdots \\ pc_{S_1^m 1} & \cdots & pc_{S_1^m p} \\ \vdots & \ddots & \vdots \\ pc_{S_{nm}^m 1} & \cdots & pc_{S_{nm}^m p} \end{pmatrix} \tag{1.18}$$

où $pc_{S_k^i j} = \sum_{l=1}^{q} x_{S_k^i l}\, u_{lj}$ est la coordonnée du sommet S_k de l'objet H_i sur l'axe factoriel de direction u_j. La position d'un objet H_i sur le $j^{ème}$ axe factoriel est définie par l'ensemble des coordonnées factorielles $\{pc_{S_1^i j}, \ldots, pc_{S_{n_i}^i j}\}$ des n_i sommets $\{S_1^i, \ldots, S_{n_i}^i\}$ de H_i cet axe. On résume l'ensemble des coordonnées des sommets d'un objet H_i sur l'axe j par l'intervalle de variation $[\underline{pc_{ij}}, \overline{pc_{ij}}]$ défini comme suit :

$$\underline{pc_{ij}} \quad = \quad \min_{(k=1..n_i)} pc_{S_k^i j} \tag{1.19}$$

$$\overline{pc_{ij}} \quad = \quad \max_{(k=1..n_i)} pc_{S_k^i j} \tag{1.20}$$

$[\underline{pc_{ij}}, \overline{pc_{ij}}]$ constitue la coordonnée factorielle (de type intervalle) de l'objet H_i sur l'axe factoriel j.

1.2.7 La projection d'un point de l'hyper-rectangle H_i

Un objet H_i est représenté dans l'espace factoriel par un hyper-rectangle dont la description est :

$$H_i \;=\; \Big([\underline{pc_{i1}}, \overline{pc_{i1}}], \ldots, [\underline{pc_{iq}}, \overline{pc_{iq}}]\Big)$$

Soit $\tilde{x_i} = (\tilde{x_{i1}}, \ldots, \tilde{x_{iq}})$ avec $\tilde{x_{ij}} \in [\underline{x_{ij}}, \overline{x_{ij}}]$ un point quelconque de l'hyper-rectangle H_i. Montrons que la projection, dans l'espace factoriel, d'un point quelconque $\tilde{x_i}$ de l'hyper-rectangle H_i est contenue dans la projection factorielle de l'hyper-rectangle H_i. Pour cela, montrons que la coordonnée factorielle $\tilde{pc_{ij}}$ du point $\tilde{x_i}$ sur l'axe factoriel de direction u_j est une valeur comprise dans l'intervalle $[\underline{pc_{ij}}, \overline{pc_{ij}}]$. Par définition, la coordonnée factorielle $\tilde{pc_{ij}}$ s'exprime comme suit :

$$\tilde{pc_{ij}} \;=\; \sum_{l=1}^{q} \tilde{x_{il}} u_{lj} \tag{1.21}$$

Sachant que chaque valeur $\tilde{x_{il}}$ varie dans $[\underline{x_{il}}, \overline{x_{il}}]$, on a :

$$\begin{array}{rcll}
\underline{x_{ij}} \leq & \tilde{x_{il}} & \leq \overline{x_{ij}} & \\
\underline{x_{ij}}\ u_{lj} \leq & \tilde{x_{il}}\ u_{lj} & \leq \overline{x_{ij}}\ u_{lj} & \text{si } \ u_{lj} > 0 \\
\underline{x_{ij}}\ u_{lj} \geq & \tilde{x_{il}}\ u_{lj} & \geq \overline{x_{ij}}\ u_{lj} & \text{si } \ u_{lj} < 0
\end{array}$$

ainsi, $\tilde{pc_{ij}}$ est borné comme suit :

$$\begin{aligned}
\tilde{pc_{ij}} \;&=\; \sum_{l=1}^{q} \tilde{x_{il}} u_{lj} = \sum_{(l=1\ /\ u_{lj}>0)}^{q} \tilde{x_{il}}\ u_{lj} \;+\; \sum_{(l=1\ /\ u_{lj}<0)}^{q} \tilde{x_{il}}\ u_{lj} \\
&\geq \sum_{(l=1\ /\ u_{lj}>0)}^{q} \underline{x_{il}}\ u_{lj} \;+\; \sum_{(l=1\ /\ u_{lj}<0)}^{q} \overline{x_{il}}\ u_{lj} \\
&\leq \sum_{(l=1\ /\ u_{lj}<0)}^{q} \underline{x_{il}}\ u_{lj} \;+\; \sum_{(l=1\ /\ u_{lj}>0)}^{q} \overline{x_{il}}\ u_{lj}
\end{aligned}$$

d'autre part,

$$\underline{pc_{ij}} \;=\; \min_{(k=1..n_i)} pc_{S_k^i j} = \min_{(k=1..n_i)} \sum_{l=1}^{q} x_{S_k^i l}\ u_{lj}$$

comme les variables $x_{S_k^i l}$ $(l = 1..q)$ sont supposées indépendantes (indépendance locale), alors :

$$\underline{pc_{ij}} = \sum_{l=1}^{q} \min_{(k=1..n_i)} x_{S_k^i l}\, u_{lj} \tag{1.22}$$

$$= \sum_{(l=1\ /\ u_{lj}>0)}^{q} \min_{(k=1..n_i)} x_{S_k^i l}\, u_{lj} + \sum_{(l=1\ /\ u_{lj}<0)}^{q} \min_{(k=1..n_i)} x_{S_k^i l}\, u_{lj} \tag{1.23}$$

$$= \sum_{(l=1\ /\ u_{lj}>0)}^{q} \underline{x_{il}}\, u_{lj} + \sum_{(l=1\ /\ u_{lj}<0)}^{q} \overline{x_{il}}\, u_{lj} \tag{1.24}$$

de manière similaire on a :

$$\overline{pc_{ij}} = \sum_{(l=1\ /\ u_{lj}<0)}^{q} \underline{x_{il}}\, u_{lj} + \sum_{(l=1\ /\ u_{lj}>0)}^{q} \overline{x_{il}}\, u_{lj} \tag{1.25}$$

ainsi, $\forall\ \tilde{x_{ij}} \in [\underline{x_{ij}}, \overline{x_{ij}}]\ \ \tilde{pc_{ij}} \in [\underline{pc_{ij}}, \overline{pc_{ij}}]$.

1.2.8 Le nuage des variables de type intervalle

À toute variable X_j de type intervalle, on fait correspondre un vecteur $y_j \in \Re^n$ muni de la masse unité :

$$y_j = \begin{pmatrix} x_{S_1^1 j} \\ \vdots \\ x_{S_{n_1}^1 j} \\ \vdots \\ x_{S_1^m j} \\ \vdots \\ x_{S_{n_m}^m j} \end{pmatrix} \tag{1.26}$$

Le choix de la distance dans $\Re^n$ consiste à affecter à chaque dimension un coefficient égal au poids de chaque objet dans le nuage $N(H)$ de $\Re^q$. Ainsi, l'espace de description des variables est muni de la métrique D définie en 1.7.

On note G_j la $j^{ème}$ composante principale donnant la description des variables sur le $j^{ème}$ axe factoriel ; elle est définie comme suit :

$$G_j = \sqrt{\lambda_j} u_j \tag{1.27}$$

1.2.9 Représentation graphique des objets H_i

Soit la matrice suivante à m lignes et p colonnes donnant la description des objets H_i par les p premières composantes principales de type intervalle :

$$\begin{pmatrix} H_1 \\ \vdots \\ H_m \end{pmatrix} = \begin{pmatrix} [\underline{pc_{11}}, \overline{p_{11}}] & \cdots & [\underline{pc_{1p}}, \overline{pc_{1p}}] \\ \vdots & \ddots & \vdots \\ [\underline{pc_{m1}}, \overline{pc_{m1}}] & \cdots & [\underline{pc_{mp}}, \overline{pc_{mp}}] \end{pmatrix} \tag{1.28}$$

Les objets H_i décrits par des composantes principales de type intervalle sont visualisés dans les plans factoriels par des points, des segments ou des rectangles. On utilisera dans ce qui suit le terme rectangle pour également désigner aussi bien un segment qu'un point. La figure 1.4 suivante montre la projection d'un objet H_i (décrit initialement par trois variables X_1, X_2, X_3 de type intervalle) dans l'espace factoriel à trois dimensions issues de l'analyse en composantes principales d'un nuage d'objets.

Comme le montre la figure précédente, la représentation de la projection d'un hyper-rectangle, dans un plan factoriel, par un rectangle, constitue une enveloppe maximale : tout point à l'intérieur de l'hyper-rectangle est projeté dans le rectangle (parties non hachurées) mais tout point du rectangle n'est pas forcement la projection d'un point de l'hyper-rectangle (parties hachurées). La description la plus précise d'un objet, dans un plan factoriel, est donnée par la couverture convexe de la projection de l'objet sur ce plan.

On adopte plutôt une représentation rectangulaire pour plusieurs raisons. D'une part, les couvertures convexes nécessitent souvent l'utilisation de fonctions complexes difficilement interprétables par les utilisateurs ; d'autre part, la conservation du principe d'une ACP qui consite à réduire la dimension de

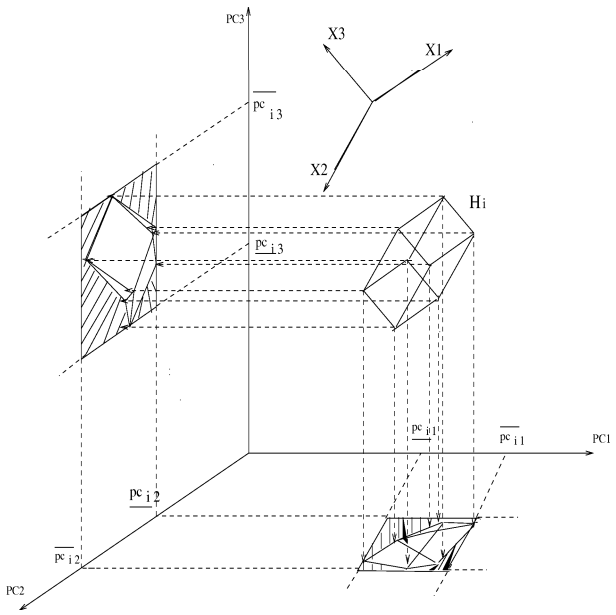


FIG. 1.4:

l'espace de description tout en conservant le type des variables d'origines (revient dans ce cas à fournir des composantes principales également de type intervalle). La dernière raison est une question de portabilité, les méthodes proposées seront utilisées en amont d'autres méthodes en analyse de données symboliques, elles doivent donc fournir en sortie des descriptions utilisables par les autres méthodes.

1.2.10 Généralisation des paramètres d'aide à l'interprétation

On ne peut interpréter les proximités entre des points projetés dans un plan factoriel qu'après avoir évalué leurs paramètres d'aide à l'interprétation. On propose une généralisation des paramètres d'aide à l'interprétation usuels aux hyper-rectangles. On note $Cor(H_i, u_j)$ la contribution relative de l'objet H_i à l'axe j mesurée par son cosinus carré par rapport à cet axe. Le cosinus carré d'un objet par rapport à un espace donné (axe, plan, etc.) est une mesure de la qualité de représentation de cet objet sur l'espace en question. On propose les deux formules suivantes pour mesurer la qualité de représentation de l'objet H_i sur l'axe factoriel j :

$$Cor(H_i, u_j) = \frac{\sum_{k=1}^{n_i} p_{S_k^i} pc^2_{S_k^i j}}{\sum_{k=1}^{n_i} p_{S_k^i} d^2(S_k^i, G)} \tag{1.29}$$

$$Cor(H_i, u_j) = \frac{1}{p_i} . \sum_{k=1}^{n_i} \frac{p_{S_k^i} pc^2_{S_k^i j}}{d^2(S_k^i, G)} \tag{1.30}$$

où $p_{S_k^i}$ désignant toujours le poids du sommet S_k associé à l'objet H_i, $pc_{S_k^i j}$ la coordonnée du sommet S_k sur l'axe factoriel de direction u_j, G le centre de gravité et $d(S_k, G)$ la distance entre le sommet S_k et le centre de gravité G.

La première formule exprime le rapport entre la contribution de l'ensemble des sommets $\{S_1^i, \ldots, S_{n_i}^i\}$ à l'inertie λ_j de l'axe factoriel j et la contribution de l'ensemble des sommets à l'inertie totale, tandis que la seconde correspond à la moyenne pondérée des cosinus carrés des angles entre chacun des n_i sommets S_k^i associés à H_i et l'axe factoriel j.

Sachant que pour tout a, b, c, d réels positifs la propriété $\frac{a+c}{b+d} \leq \frac{a}{b} + \frac{c}{d}$ est toujours vérifiée, alors la première formule fournit des contributions relatives plus faibles que celles obtenues par la seconde. On mesure de même la contribution de H_i :

- à l'inertie λ_j du $j^{ème}$ axe factoriel par :

$$Ctr(H_i, u_j) = \frac{\sum_{k=1}^{n_i} p_{S_k^i} pc^2_{S_k^i j}}{\lambda_j} \tag{1.31}$$

- à l'inertie totale du nuage des n sommets associés aux m objets par :

$$Inr(H_i) = \frac{\sum_{k=1}^{n_i} p_{S_k^i} . d^2(S_k, G)}{I_T} = \frac{\sum_{k=1}^{n_i} p_{S_k^i} d^2(S_k, G)}{\sum_{j=1}^{q} \lambda_j} \tag{1.32}$$

I_T désignant l'inertie totale. Les deux contributions précédentes reviennent à effectuer des moyennes pondérées des contributions des n_i sommets associés à l'objet H_i.

1.2.11 Une visualisation qui aide à l'interprétation

Quand le nombre et le volume des objets traités est important, la visualisation rectangulaire fournit des plans factoriels encombrés et difficilement interprétables. Pour faire face à cela, on propose une procédure itérative qui permet, lors de la construction des enveloppes rectangulaires représentant les objets, de tenir compte des contributions relatives des sommets.

La procédure proposée fournit, pour chaque couple de composantes principales une succession de plans factoriels chacun associé à une qualité de représentation α ($0 \leq \alpha \leq 1$). Dans un plan factoriel de niveau α, un objet H_i est représenté par un rectangle construit en ne tenant compte que des sommets dont la qualité de représentation est supérieure ou égale à α. La $j^{ème}$ coordonnée $[\underline{pc}_{ij}^{\alpha}, \overline{pc}_{ij}^{\alpha}]$ de l'objet H_i, dans le plan factoriel de niveau α, est définie comme suit :

$$\underline{pc}_{ij}^{\alpha} = min_{k=1}^{n_i}\{pc_{S_k^i j}/Cor(S_k^i, j) \geq \alpha\} \quad (1.33)$$

$$\overline{pc}_{ij}^{\alpha} = max_{k=1}^{n_i}\{pc_{S_k^i j}/Cor(S_k^i, j) \geq \alpha\} \quad (1.34)$$

À un niveau $\alpha = 0$, on retrouve la représentation élémentaire des objets par des rectangles construits en tenant compte de tous les sommets. Si pour un objet donné, aucun sommet n'a une qualité de représentation supérieure ou égale au seuil α, l'objet est alors représenté soit par son centre étiqueté de la contribution relative de l'objet (dans le cas où l'on privilégie la visualisation des positions centrales des objets), soit par le sommet ayant la plus grande qualité de représentation (dans le cas où l'on privilégie la visualisation des sommets les plus représentatifs des objets). Chaque rectangle peut être étiqueté par l'intervalle de variation des qualités de représentation des sommets ayant servi à le construire. Les figures 1.5, 1.6 et 1.7 donnent, respectivement, la description des objets H_i et H_j dans les premiers plans factoriels de niveau zéro, de niveau 0.2 et de niveau 0.5.

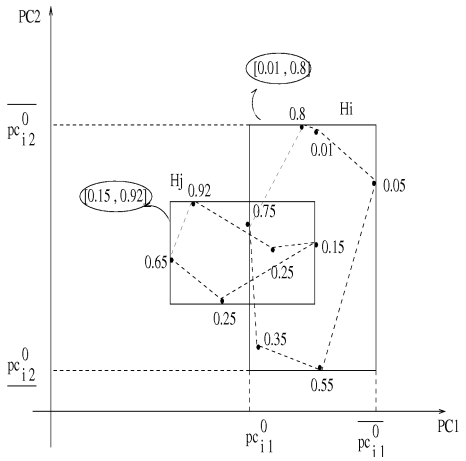


FIG. 1.5: *Projection des objets Hi et Hj dans le plan factoriel de niveau zéro*

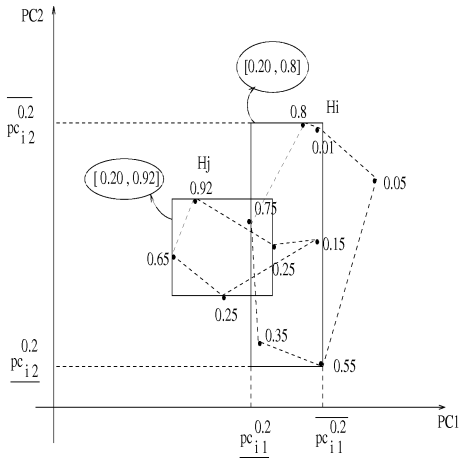


FIG. 1.6: *Projection des objets Hi et Hj dans le plan factoriel de niveau 0.2*

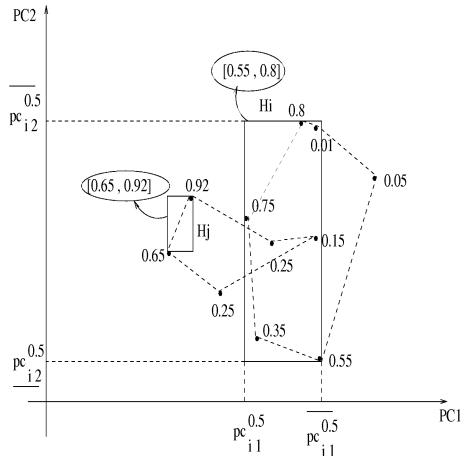


FIG. 1.7: *Projection des objets Hi et Hj dans le plan factoriel de niveau 0.5*

On constate que les objets H_i et H_j, qui se recouvrent dans le plan de niveau zéro (en considérant tous leurs sommets), sont disjoints à un niveau de qualité de représentation supérieur à 0.5. Ainsi, le recouvrement des objets à un niveau α élevé est représentatif avec un degré de confiance α du recouvrement de ces objets dans l'espace de description.

1.2.12 Algorithme de la procédure itérative de visualisation

On décrit ci-dessous la procédure permettant de visualiser les objets H_i dans le plan factoriel de niveau α défini par les axes factoriels j et l. Elle reçoit en entrée le niveau α (seuil des contributions relatives), ainsi que les coordonnées factorielles des sommets sur les axes j et l. En sortie, la procédure fournit la visualisation des objets dans le plan factoriel défini par (u_j, u_l) de niveau α. Les commentaires sont encadrés du symbole %. La procédure pourra être rappelée pour chaque niveau de qualité de représentation souhaité. On note $Cor(S_k^i, u_j, u_l)$ la contribution relative du sommet S_k de l'objet H_i dans le plan factoriel défini par les axes de directions u_j, u_l ; soit :

$$Cor(S_k^i, u_j, u_l) \quad = \quad Cor(S_k^i, u_j) + Cor(S_k^i, u_l)$$

Les fonctions $min(a, b)$ et $max(a, b)$ renvoient, respectivement, la plus petite et la plus grande des deux valeurs a et b.

Procédure Visualisation (α , PC_j, PC_l)

Pour i allant de 1 à m **Faire**

% pour chaque objet H_i %

Pour k allant de 1 à n_i **Faire**

% pour chaque sommet S_k^i de H_i %

Si ($Cor(S_k^i, u_j, u_l) \geq \alpha$) **Alors**

$\underline{pc_{ij}} = min(pc_{S_k^i j}, \underline{pc_{ij}})$

$\underline{pc_{il}} = min(pc_{S_k^i l}, \underline{pc_{il}})$

$\overline{pc_{ij}} = max(pc_{S_k^i j}, \overline{pc_{ij}})$

$\overline{pc_{il}} = max(pc_{S_k^i l}, \overline{pc_{il}})$

Fin Si

Fin Pour

Si $\forall\ S_k^i,\ \ (Cor(S_k^i, u_j, u_l) < \alpha)$

% tous les sommets de l'objet H_i ont une contribution relative inférieur à α %

Alors

$\underline{pc_{ij}} = \overline{pc_{ij}} = pc_{S_r^i j}$

$\underline{pc_{il}} = \overline{pc_{il}} = pc_{S_r^i l}$

% où S_r^i est soit le sommet de H_i doté de la plus grande contribution relative soit son centre %

Fin Si

Fin Pour

- Visualiser les objets dans le plan factoriel de niveau α

Fin procédure Visualisation

1.2.13 Algorithme de la méthode des sommets

La procédure méthode-sommets reçoit deux paramètres en entrée. Le premier est le tableau $Tab\alpha$ contenant les différents seuils de contribution relative α utilisés par la procédure de visualisation. On note *nbseuil* le nombre total de seuils α utilisés, dans les applications on utilise $nbseuil = 3$ (α=0, α=0.20, α= 0.5). Le deuxième paramètre est la matrice X_H de description des objets par les variables de type intervalle. La procédure fournit en sortie la description factorielle des objets H_i, dans les plans factoriels de niveau α. Pour chaque couple de composantes principales, on a *nbseuil* plans factoriels générés.

Procédure méthode-sommets (Tabα, X_H)

- Choix du système de pondération des objets et des sommets
- Calculer la matrice de variance-covariance V^s de la matrice X
- Diagonaliser V^s, soit F la matrice des coordonnées factorielles des sommets obtenue
- Calculer les paramètres d'aides à l'interprétation des sommets
- Généraliser les paramètres d'aide à l'interprétation aux objets H_i

Pour seuil allant de 1 à nbseuil **Faire**

% pour chaque seuil α de contribution relative dans Tabα %

Pour chaque couple PC_j, PC_l dans F **Faire**

- Visualisation(Tabα[seuil], PC_j, PC_l)

Fin Pour

Fin Pour

Fin procédure méthode-sommets

1.2.14 L'ACP, un cas particulier de la méthode des sommets

Considérons le cas particulier où tous les intervalles $[\underline{x_{ij}}, \overline{x_{ij}}]$ sont triviaux ($\underline{x_{ij}} = \overline{x_{ij}} = x_{ij}$). La matrice X_H donne alors la description des m objets par q variables numériques, soit :

$$X_H = \begin{pmatrix} H_1 \\ \vdots \\ H_m \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mq} \end{pmatrix}$$

Chaque objet H_i est alors décrit par $n_i = 2^{q_i} = 2^0 = 1$ sommet. On note S^i l'unique sommet représentant l'objet H_i. La première étape dans la méthode des sommets consiste à décrire chaque hyper-rectangle H_i par l'ensemble de ses sommets. Chaque objet H_i est alors décrit par une matrice X_{H_i} à n_i lignes donnant la description de ses n_i sommets dans l'espace de description. Dans le cas présent, un objet H_i est décrit par un vecteur ligne X_{H_i} à $n_i = 1$ ligne et

q colonnes donnant la description du sommet S^i dans l'espace de description. Le vecteur ligne X_{H_i} correspond à la *ème* ligne $(x_{i1}, \ldots, x_{iq})$ dans la matrice de départ X_H. Par conséquent, la matrice de données X (concaténation des X_{H_i}) à analyser dans la méthode des sommets est exactement la matrice de départ X_H analysée dans le cas d'une ACP classique. L'étape suivante consiste à diagonaliser la matrice de variance-covariance associée à $X = X_H$. Dans la méthode des sommets, les paramètres d'aide à l'interprétation d'un objet sont les moyennes pondérées (seconde formule pour Cor) des paramètres d'aide à l'interprétation de ses sommets. Dans ce cas présent, les paramètres d'aide à l'interprétation de l'objet H_i sont ceux du seul point représentatif S^i, comme dans le cas d'une ACP classique.

Dans la procédure de visualisation itérative, un objet H_i décrit par un seul sommet ($q_i = 0$) est représenté dans un plan factoriel de niveau α par un point, dont la position reste inchangée quelle que soit la valeur de α. En effet, dans le cas où le sommet représentant l'objet a une contribution relative supérieure à α, il est alors le point représentatif de l'objet dans le plan factoriel de niveau α ; dans le cas contraire, l'objet est représenté par le sommet ayant la plus grande contribution relative, c'est-à-dire l'unique sommet en question. Dans ce cas particulier, où tous les objets sont décrits par des points, la procédure de visualisation fournit, pour chaque couple de direction (u_j, u_l) et pour n'importe quelle valeur α, un même plan factoriel identique à celui fourni par une ACP classique. La méthode des sommets est donc bien une généralisation de l'ACP classique.

1.3 Cas de données intervalles dotés de contraintes de domaines

1.3.1 Contraintes de domaines associées aux objets

Dans ce qui précéde, on considère qu'un objet H_i définit une région de valeurs dans l'espace de description et que tout point à l'intérieur de l'hyper-rectangle constitue une observation possible du monde. En pratique, on peut être confronté à la situation où seulement un sous-domaine de l'hyper-rectangle est observable (valide) ; c'est dans le cas où des contraintes de domaines sont définies entre les variables. Considérons les deux exemples suivants, illustrant le cas de variables intervalles liées par des contraintes de domaines.

1- Exemple : Soit la description du quotient familial QF et de l'impôt sur le revenu en (en KF) d'une population d'individus P_i :

$$P_i \;=\; [\text{QF} \in [60, 200]] \;\wedge\; \text{Impôt} \in [4.806, 53.821]$$

La contrainte de domaine liant les deux variables est décrite par les trois règles suivantes :

a) Si QF $\in [60, 89.65]$, alors Impôt $\in [4.806, 11.922]$,

b) Si QF $\in [89.65, 145.16]$, alors Impôt $\in [11.922, 30.24]$,

c) Si QF $\in [145.16, 200]$, alors Impôt $\in [30.24, 53.821]$.

La visualisation du domaine de variation de la population P_i ainsi que des régions des valeurs observables (valides) ou non-observables (non-valides) est :

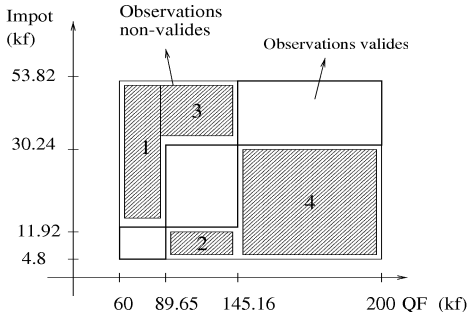


Fig. 1.8:

La description des quatre régions non-valides dans l'espace des deux variables est :

$$\begin{aligned}
\text{Région 1} &= ([60, 89.65], [11.92, 53.82]) \\
\text{Région 2} &= ([89.65, 145.16], [4.8, 11.92]) \\
\text{Région 3} &= ([89.65, 145.16], [30.24, 53.82]) \\
\text{Région 4} &= ([145.16, 200], [4.8, 30.24])
\end{aligned}$$

2- Exemple : Soit la description probabiliste de l'objet H_i par deux variables :

$$H_i = [X_1 \in [0.2, 0.7]] \wedge [X_2 \in [0.1, 0.6]]$$

La contrainte de domaine liant les valeurs prises par X_1 et X_2 est :

$$\forall\ x_{i1} \in [0.2, 0.7]\ \ \forall x_{i2} \in [0.1, 0.6] \implies x_{i1} + x_{i2} = 1$$

Le domaine des régions valides est un segment de la droite $X_2 = -X_1 + 1$:

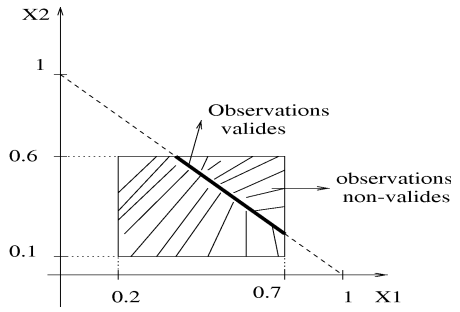


FIG. 1.9:

De manière plus générale, les valeurs valides peuvent être définies par une fonction à valeurs dans le domaine et de paramètres les variables liées par les contraintes.

On s'intéresse ici à des contraintes de domaine du type défini dans l'exemple 1. Sachant qu'un hyper-rectangle ne délimite pas forcément une région homogène mais peut inclure une ou plusieurs régions non-valides, notre objectif consiste à étendre les méthodes précédemment proposées pour qu'elle puisse tenir compte d'une telle information.

Une contrainte de domaine, du type défini dans l'exemple 1, peut être caractérisée par la description de l'ensemble de ses régions valides ou par la description de l'ensemble de ses régions non-valides.

On note $C^i_1, \ldots, C^i_{r_i}$ les r_i régions non-valides disjointes associées à l'hyper-rectangle H_i et définies comme suit :

$\forall\ l = 1..r_i\ \ C^i_l = \wedge_{(j \in E_{C^i_l})}\ [\underline{c^i_{lj}}, \overline{c^i_{lj}}]$ avec, $\forall\ j \in E_{C^i_l}$, alors $[\underline{c^i_{lj}}, \overline{c^i_{lj}}] \subset [\underline{x_{ij}}, \overline{x_{ij}}]$.

$E_{C_l^i} \subset \{1, .., q\}$ est l'ensemble des indices des variables $X_1, ..., X_q$ impliquées dans la description de C_l^i.

Une région non-valide C_l^i est représentée par un hyper-rectangle (dit hyper-rectangle contrainte) inclus dans l'hyper-rectangle H_i. On note $S_1^{C_l^i}, ..., S_{n_{il}}^{C_l^i}$ les n_{il} sommets associés à l'hyper-rectangle contrainte C_l^i. On suppose, dans ce qui suit, que les hyper-rectangles contraintes sont disjoints et que les règles les définissant sont cohérentes entre elles.

Soit, par exemple, la visualisation dans la figure 1.10 de l'objet $H_i = ([\underline{x_{i1}}, \overline{x_{i1}}], [\underline{x_{i2}}, \overline{x_{i2}}], [\underline{x_{i3}}, \overline{x_{i3}}])$ décrit par trois variables de type intervalle :

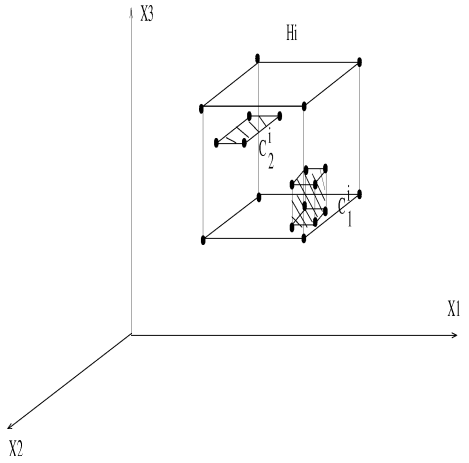


FIG. 1.10:

C_1^i, C_2^i deux hyper-rectangles contraintes décrits comme suit :

$$\begin{aligned} C_1^i &= \left([\underline{c_{11}^i}, \overline{c_{11}^i}], [\underline{c_{12}^i}, \overline{c_{12}^i}], [\underline{c_{13}^i}, \overline{c_{13}^i}]\right) \\ C_2^i &= \left([\underline{c_{21}^i}, \overline{c_{21}^i}], [\underline{c_{22}^i}, \overline{c_{22}^i}], c_{23}^i\right) \end{aligned}$$

1.3.2 Pondération sous-contraintes des objets H_i

Étant donné un ensemble d'objets H_i dotés d'un ensemble de contraintes de domaine C_l^i $(l = 1..r_i)$, notre objectif est de tenir compte de cette information

dans la méthode des sommets et dans la méthode des centres. Rappelons que chaque objet H_i de poids p_i est représenté par n_i sommets $S_1, ..., S_{n_i}$. Dans la méthode des sommets, le poids p_i est réparti sur l'ensemble des sommets de l'objet H_i ; on note $p_{S_k^i}$ le poids du sommet S_k de l'objet H_i.

Pour tenir compte des contraintes de domaine on propose, selon la stratégie de pondération adoptée proportionnelle aux volumes ou inversement proportionnelle aux volumes, de décroître, respectivement, croître le poids d'un objet H_i proportionnellement aux volumes des hyper-rectangles contraintes qui lui sont associés. On note $V(C_l^i)$ le volume de l'hyper-rectangle contrainte C_l^i associé à H_i, défini comme suit :

$$V(C_l^i) = \prod_{(j \in E_{C_l^i})} (\overline{c_{kj}^i} - \underline{c_{kj}^i})$$

et $V(H_i)$ le volume total de l'hyper-rectangle H_i :

$$V(H_i) = \prod_{(\overline{x_{ij}} \neq \underline{x_{ij}})} (\overline{x_{ij}} - \underline{x_{ij}})$$

On définit le nouveau volume $V(H_i)^\star$ de H_i tenant compte des contraintes de domaine $C_1^i, ..., C_{r_i}^i$, comme suit :

$$V(H_i)^\star = V(H_i) - \sum_{s=1}^{r_i} V(C_s^i)$$

On note $p_1^\star, \ldots, p_m^\star$ les nouveaux poids des objets calculés en fonction des nouveaux volumes $V(H_1)^\star, \ldots, V(H_m)^\star$ et selon la stratégie de pondération adoptée dans la méthode.

1.3.3 Pondération sous-contraintes des sommets S_k^i

La pondération sous-contraintes des sommets intervient uniquement dans le cas de la méthode des sommets. Il s'agit de calculer les poids des sommets sur la base des nouveaux poids $p_i^\star$. Le principe est simple, plus un sommet est proche des régions non-valides plus faible est son poids. Pour ce faire, on propose de répartir le poids $p_i^\star$ de H_i sur l'ensemble des sommets en tenant compte de la position de chaque sommet S_k^i par rapport à l'ensemble des

hyper-rectangles contraintes C_l^i. Pour cela, on calcule pour chaque sommet S_k^i sa distance par rapport à chaque hyper-rectangle contrainte C_l^i. On définit la distance entre un sommet S_k^i et un hyper-rectangle contrainte comme la moyenne des distances entre S_k^i et l'ensemble des sommets de C_l^i, soit :

$$D(S_k^i, C_l^i) = \frac{1}{n_{il}} \sum_{l=1}^{n_{il}} d(S_k^i, S_l^{C_l^i})$$

où d est la distance euclidienne entre deux points sommets, et n_{il} le nombre de sommets de l'hyper-rectangle contrainte C_l^i. On définit alors la distance moyenne d'un sommet S_k^i à l'ensemble des hyper-rectangles contraintes C_l^i comme suit :

$$d(S_k^i) = \frac{1}{r_i} \sum_{l=1}^{r_i} D(S_k^i, C_l^i)$$

$p^\star_{S_k^i}$ le nouveau poids du sommet S_k^i tenant compte des contraintes de domaine de l'objet H_i est défini comme suit :

$$p^\star_{S_k^i} = p_i^\star \frac{d(S_k^i)}{\sum_{k=1}^{n_i} d(S_k^i)}$$

On vérifie aisément que $\sum_{k=1}^{n_i} p^\star_{S_k^i} = p_i^\star$.

1.4 La méthode des centres

1.4.1 Les données

On note x_i^o le point de référence de l'hyper-rectangle H_i considéré comme le point le plus représentatif de l'objet :

$$x_i^o = \left(x_{i1}^o, \ldots, x_{iq}^o\right) \tag{1.35}$$

où x_{ij}^o la coordonnée du barycentre sur l'axe X_j est définie comme suit :

$$x_{ij}^o = \sum_{k=1}^{n_i} \frac{p_{S_k^i}}{p_i} x_{S_k^i j} \tag{1.36}$$

D'après ce qui a été établi en 1.6, x_{ij}^o s'écrit :

$$x_{ij}^o = (\underline{p_{ij}x_{ij}} + \overline{p_{ij}x_{ij}})$$

Soit X la matrice à m lignes et q colonnes suivante :

$$X = \begin{pmatrix} x_{11}^o & \cdots & x_{1q}^o \\ \vdots & \ddots & \vdots \\ x_{m1}^o & \cdots & x_{mq}^o \end{pmatrix} \tag{1.37}$$

1.4.2 Le nuage d'objets N(H)

À tout objet $H_i \in N(H)$, on fait correspondre un vecteur x_i^o de $\Re^q$ tel que :

$$x_i^o = \left(x_{i1}^o, \ldots, x_{iq}^o\right) \tag{1.38}$$

Le nuage $N(H)$ des centres des objets H_i, décrits par la matrice 1.1, est représenté dans la figure 1.11 suivante :

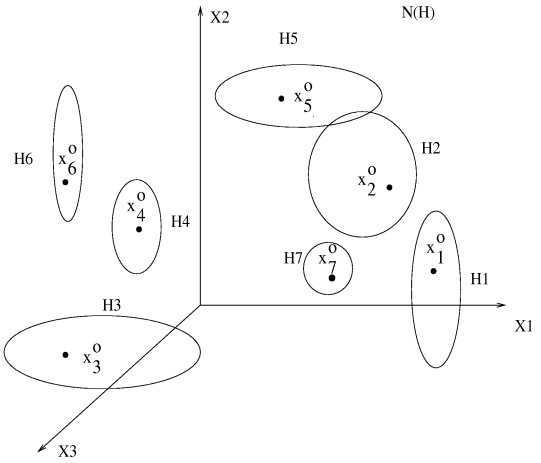


FIG. 1.11: *Nuage N(H) des point de références des hyper-rectangles*

L'ensemble des objets H_i est défini dans l'espace $\Re^q$, par la matrice X muni de la métrique identité.

1.4.3 Pondération des objets et sémantiques sous-jacentes

Contrairement à la méthode des sommets où le poids p_i de l'objet H_i est réparti entre ses sommets, dans la méthode des centres on suppose que le poids p_i est accumulé au point de référence x_i^o. De manière similaire, on peut adopter un système de pondération tenant compte de la sémantique des intervalles tel que c'est défini dans la méthode des sommets (paragraphes 1.2.3 et 1.3.2). Soit D la matrice diagonale des poids des points de références à m lignes et q colonnes définie comme suit :

$$D = \begin{pmatrix} p_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p_m \end{pmatrix} \tag{1.39}$$

1.4.4 Matrice de variance-covariance

Supposons, ce qui ne restreint pas la généralité, que les variables X_j sont centrées. La matrice de variance-covariance V^c de la matrice X est définie comme suit :

$$V^c = X^T D X \tag{1.40}$$

La moyenne dans la méthode des centre $\overline{X_j^c}$ de X_j s'écrira donc :

$$\overline{X_j^c} = \sum_{i=1}^{m} p_i x_{ij}^o = 0$$

La variance v_{jj}^c de la variable Xj dans la méthode des centres s'écrit :

$$\begin{aligned} v_{jj}^c &= \sum_{i=1}^{m} p_i (x_{ij}^o)^2 \\ &= \sum_{i=1}^{m} p_i (\underline{p_{ij}}\,\underline{x_{ij}} + \overline{p_{ij}}\,\overline{x_{ij}})^2 \end{aligned}$$

Le terme général de la covariance v_{jl}^c entre les variables X_j et X_l s'écrit :

$$v^c_{jl} = \sum_{i=1}^{n} p_i x^o_{ij} x^o_{il} \tag{1.41}$$

$$= \sum_{i=1}^{n} p_i(\underline{p_{ij}}\,\underline{x_{ij}} + \overline{p_{ij}}\,\overline{x_{ij}})(\underline{p_{il}}\,\underline{x_{il}} + \overline{p_{il}}\,\overline{x_{il}}) \tag{1.42}$$

1.4.5 Description factorielle du nuage d'objets

La méthode des centres se fonde, pour la détermination des axes factoriels d'inertie, sur l'information apportée par les centres (points de références) des hyper-rectangles. La matrice à diagonaliser est V^c. Soient u_j, λ_j et PC_j, respectivement, le $j^{ème}$ vecteur propre, la $j^{ème}$ valeur propre associée et la $j^{ème}$ composante principale.

$$PC_j = X.u_j$$

On note F la matrice à m lignes et p colonnes donnant la description factorielle des m centres des hyper-rectangles dans l'espace défini par les p premières composantes principales $PC_1, \ldots, PC_p$:

$$F = \begin{pmatrix} pc_{x^o_1 1} & \cdots & pc_{x^o_1 p} \\ \vdots & \ddots & \vdots \\ pc_{x^o_m 1} & \cdots & pc_{x^o_m p} \end{pmatrix} \tag{1.43}$$

où $pc_{x^o_i j}$ est la coordonnée du centre x^o_i de l'hyper-rectangle H_i sur l'axe factoriel j. Les positions des centres des hyper-rectangles dans le plan factoriel fournissent une tendance centrale de la distribution des objets.

Pour restituer lors de la description factorielle l'information de variation ou d'imprécision intrinsèque à chaque objet, on fait varier chaque centre au sein de son hyper-rectangle et on déduit les intervalles de variation de ce point sur chaque axe factoriel. Soit $x_i = (x_{i1}, \ldots, x_{iq})$ un point quelconque de l'hyper-rectangle H_i. La coordonnée du point x_i sur l'axe factoriel j est donnée par la fonction suivante :

$$pc_{x_i j} = \sum_{l=1}^{q} x_{il}\, u_{lj} \tag{1.44}$$

Quand on fait varier le point x_i à l'intérieur de l'hyper-rectangle, la plus petite coordonnée $\underline{pc_{ij}}$ et la plus grande coordonnée $\overline{pc_{ij}}$ prises par le point x_i sur l'axe factoriel j sont obtenues en minimisant et en maximisant la fonction définie en 1.44, soit :

$$\begin{aligned} \underline{pc_{ij}} &= \min_{(\underline{x_{il}} \leq x_{il} \leq \overline{x_{il}})} \left(\sum_{l=1}^{q} x_{il}\, u_{lj} \right) \\ \overline{pc_{ij}} &= \max_{(\underline{x_{il}} \leq x_{il} \leq \overline{x_{il}})} \left(\sum_{l=1}^{q} x_{il}\, u_{lj} \right) \end{aligned}$$

d'après les résultats établis dans la section 1.2.7 on a :

$$\underline{pc_{ij}} = \sum_{(l=1,u_{lj}<0)}^{q} \overline{x_{il}}\, u_{lj} + \sum_{(l=1,u_{lj}>0)}^{q} \underline{x_{il}}\, u_{lj} \tag{1.45}$$

$$\overline{pc_{ij}} = \sum_{(l=1,u_{lj}<0)}^{q} \underline{x_{il}}\, u_{lj} + \sum_{(l=1,u_{lj}>0)}^{q} \overline{x_{il}}\, u_{lj} \tag{1.46}$$

ainsi, $\underline{pc_{ij}}$ et $\overline{pc_{ij}}$ correspondent, respectivement, à la plus petite et à la plus grande coordonnée factorielle des n_i sommets projetés en supplémentaire sur l'axe factoriel j.

1.4.6 Visualisation et interprétation du nuage N(H)

De manière similaire, les m objets H_i sont décrits dans l'espace factoriel par des composantes principales de type intervalle. Afin d'obtenir une représentation graphique et une interprétation plus précise des positions des hyper-rectangles, on propose d'utiliser la procédure de visualisation itérative définie dans la méthode des sommets. Pour cela, on calcule les contributions relatives des sommets (éléments supplémentaires) dans les plans factoriels.

Les centres des hyper-rectangles étant les seuls éléments actifs dans l'analyse, par conséquent, les contributions absolues et les contributions des hyper-rectangles à l'inertie sont celles des centres des hyper-rectangles. Les qualités de représentation des hyper-rectangles sont déduites à partir des qualités de représentation de leurs sommets (projetés en supplémentaires), tel que c'est défini en (1.29).

1.4.7 Algorithme de la méthode des centres

La procédure méthode-centres reçoit comme premier paramètre le tableau $Tab\alpha$ contenant les différents seuils de contribution relative α utilisés par la procédure de visualisation et comme second paramètre la matrice X_H de description des objets par les variables de type intervalle. La procédure fournit en sortie la description factorielle des objets H_i dans les plans factoriels de niveau α. Pour chaque couple de composantes principales on a *nbseuil* plans factoriels générés.

Procédure méthode-centres (Tabα, X_H)

- Choix du système de pondération des objets
- Calculer la matrice de variance-covariance V^c de la matrice X
- Diagonaliser V^c, soit F la matrice des coordonnées factorielles des centres

Pour i allant de 1 à m **Faire**

% Pour chaque objet H_i %

Pour j allant de 1 à q **Faire**

% Pour chaque axe factoriel de direction u_j %

- Calculer l'intervalle de variation $[\underline{pc_{ij}}, \overline{pc_{ij}}]$ sur chaque axe factoriel u_j.

Fin Pour

- Visualiser, dans chaque plan factoriel, le rectangle associés à l'objet H_i

% cela revient à une visualisation de H_i dans le plan factoriel de niveau 0 %

Fin Pour

- Calculer les contributions absolues et à l'inertie des centres
- Calculer les contributions relatives des sommets des hyper-rectangles en tant que éléments supplémentaires.
- Calculer les paramètres d'aide à l'interprétation associés aux objets H_i

Pour seuil allant de 1 à nbseuil **Faire**

% pour chaque seuil α de contribution relative dans Tabα %

Pour chaque couple PC_j, PC_l dans F **Faire**

- Visualisation(Tabα[seuil], PC_j, PC_l)

Fin Pour

Fin Pour

Fin procédure méthode-centres

1.4.8 L'ACP, un cas particulier de la méthode des centres

Considérons le cas particulier où tous les intervalles $[\underline{x_{ij}}, \overline{x_{ij}}]$ sont triviaux ($\underline{x_{ij}} = \overline{x_{ij}} = x_{ij}$). La matrice X_H donne alors la description des m objets par q variables numériques, soit :

$$X_H = \begin{pmatrix} H_1 \\ \vdots \\ H_m \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mq} \end{pmatrix}$$

Chaque objet H_i est décrit par $n_i = 2^{q_i} = 2^0 = 1$ sommet. On note S^i l'unique sommet représentant l'objet H_i. La première étape dans la méthode des centres consiste à décrire chaque hyper-rectangle H_i par son centre x_i^o. Dans le cas particulier où un objet H_i est représenté par un seul sommet ($q_i = 0$), le centre x_i^o d'un tel objet est le sommet S^i en question. La matrice des centres X à analyser dans la méthode des centres est donc la matrice de données de départ X_H. La seconde étape consiste à diagonaliser la matrice de variance-covariance associée à la matrice de départ X_H, puis à projeter les centres $x_i^o = S^i$ dans les plans factoriels. L'étape suivante consiste à projeter les sommets des hyper-rectangles en supplémentaires afin de déduire la variation intrinsèque à chaque objet. Comme il n'y a aucun autre sommet que S^i, cette procédure n'est pas effectuée. Finalement, les paramètres d'aide à l'interprétation d'un objet H_i sont ceux de l'unique sommet S^i le décrivant. La méthode des centres ainsi définie est bien une ACP classique de l'ensemble des objets décrits par la matrice X_H.

1.5 Comparaison de la méthode des sommets et de la méthode des centres

Notons que dans la méthode des sommets les axes d'inerties sont déterminés à partir des sommets alors que dans la méthode des centres ils sont définis par les centres des hyper-rectangles. Dans le cas de données intervalles de faibles amplitudes (i.e. le cas des intervalles exprimant une imprécision autour d'une mesure) la méthode des centres et la méthode des sommets fournisssent des résultats très similaires. En effet, dans le cas d'intervalles de faibles amplitudes, la distribution des centres est très proche de celle des sommets des hyper-rectangles. Par conséquent, l'analyse en composantes principales du nuage des

centres et l'analyse en composantes principales du nuage des sommets fournissent des résultats (axes principaux, valeurs propres, etc.) similaires. Ces deux méthodes fournissent en revanche des résultats très différents dans le cas de données intervalles de grandes amplitudes (i.e. le cas des intervalles exprimant de la variation). En effet, la distribution du nuage des centres peut être très différente de celle des sommets des hyper-rectangles.

Nous comparons dans ce qui suit les expressions de la moyenne, variance et covariance dans le cas de ces deux méthodes et nous montrons que l'analyse en composantes principales du nuage des centres correspond à une **analyse inter-objets** , alors que celle du nuage des sommets correspond à une **analyse inter-objets** et **intra-objets**.

Les moyennes X_j^c et X_j^s

Rappelons l'expression de la moyenne d'une variable X_j dans le cas de la méthode des centres :

$$\overline{X_j^c} = \sum_{i=1}^{m} p_i x_{ij}^o$$

remplaçons x_{ij}^o par son expression définie en 1.6 :

$$\begin{aligned} \overline{X_j^c} &= \sum_{i=1}^{m} p_i(\underline{p_{ij}}\,\underline{x_{ij}} + \overline{p_{ij}}\,\overline{x_{ij}}) \\ &= \sum_{i=1}^{m}(\underline{\alpha_{ij}}\,\underline{x_{ij}} + \overline{\alpha_{ij}}\,\overline{x_{ij}}) = \overline{X_j^s} \end{aligned}$$

Ainsi, on obtient la même moyenne pour la variable X_j dans la méthode des centres et dans la méthode des sommets :

$$\forall\ j \in \{1..q\} \qquad X_j^c = X_j^s$$

Les variances V_{jj}^s et V_{jj}^c

Supposons les variables centrées et calculons la différence entre la variance de la variable X_j dans les deux méthodes :

$$
\begin{aligned}
V_{jj}^s - V_{jj}^c &= \sum_{i=1}^{m}(\underline{\alpha_{ij}}\ \underline{x_{ij}}^2 + \overline{\alpha_{ij}}\ \overline{x_{ij}}^2) - \sum_{i=1}^{m} p_i(\underline{p_{ij}}\ \underline{x_{ij}} + \overline{p_{ij}}\ \overline{x_{ij}})^2 \\
&= \sum_{i=1}^{m}(p_i\underline{p_{ij}}\ \underline{x_{ij}}^2 + p_i\overline{p_{ij}}\ \overline{x_{ij}}^2) - \sum_{i=1}^{m} p_i(\underline{p_{ij}}^2\ \underline{x_{ij}}^2 + \overline{p_{ij}}^2\ \overline{x_{ij}}^2 + 2\underline{p_{ij}}\ \overline{p_{ij}}\underline{x_{ij}}\overline{x_{ij}}) \\
&= \sum_{i=1}^{m} p_i\left(\underline{p_{ij}}\ \underline{x_{ij}}^2 + \overline{p_{ij}}\ \overline{x_{ij}}^2 - \underline{p_{ij}}^2\ \underline{x_{ij}}^2 - \overline{p_{ij}}^2\ \overline{x_{ij}}^2 - 2\underline{p_{ij}}\ \overline{p_{ij}}\underline{x_{ij}}\overline{x_{ij}}\right) \\
&= \sum_{i=1}^{m} p_i\left(\underline{p_{ij}}\ \underline{x_{ij}}^2(1-\underline{p_{ij}}) + \overline{p_{ij}}\ \overline{x_{ij}}^2(1-\overline{p_{ij}}) - 2\underline{p_{ij}}\ \overline{p_{ij}}\underline{x_{ij}}\overline{x_{ij}}\right) \\
&= \sum_{i=1}^{m} p_i\underline{p_{ij}}\ \overline{p_{ij}}\ (\overline{x_{ij}} - \underline{x_{ij}})^2
\end{aligned}
$$

où, rappelons-le, $\underline{p_{ij}} = 1 - \overline{p_{ij}}$. Ainsi, la variance V^s de la variable X_j dans le cas de la méthode des sommets est égale à la variance V^c de la variable dans le cas de la méthode des centres augmentée d'un facteur e_{jj} positif. Soit encore, $\forall\ j \in \{1..q\}\ \ v_{jj}^s = v_{jj}^c + e_{jj}$ avec,

$$
e_{jj} = \sum_{i=1}^{m} p_i\underline{p_{ij}}\ \overline{p_{ij}}\ (\overline{x_{ij}} - \underline{x_{ij}})^2
$$

Le facteur amplitude e_{jj} de la variable X_j exprime l'information de variation ou d'imprécision de la variable de type intervalle X_j. Ce facteur est nul dans le cas où la variable X_j est numérique. On note E la matrice diagonale de terme général e_{jj} :

$$
E = \begin{pmatrix} e_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & eqq \end{pmatrix} \tag{1.47}
$$

Les covariances v_{jl}^s et v_{jl}^c

Soit l'expression de la covariance v_{jl}^c entre les variables X_j et X_l $(j \neq l)$ dans la méthode des centres tel que c'est établi en 1.42 :

$$v^c_{jl} = \sum_{i=1}^{n} p_i(\underline{p_{ij}}\,\underline{x_{ij}} + \overline{p_{ij}}\,\overline{x_{ij}})(\underline{p_{il}}\,\underline{x_{il}} + \overline{p_{il}}\,\overline{x_{il}})$$

En développant le produit on obtient l' expression de la covariance établie dans le cas de la méthode des sommets.

$$\forall\ j\ , l \in \{1..q\} \qquad v^s_{jl} = v^c_{jl} \tag{1.48}$$

1.5.1 Une analyse inter-objets et intra-objets

D'après les résultats précédents on peut établir la relation suivante entre les matrices de variance-covariance V^s et V^c à diagonaliser, respectivement, dans la méthode des sommets et dans la méthode des centres.

$$V^s = \begin{pmatrix} v^s_{11} & \cdots & v^s_{1q} \\ \vdots & \ddots & \vdots \\ v^s_{q1} & \cdots & v^s_{qq} \end{pmatrix} = \begin{pmatrix} v^c_{11} & \cdots & v^c_{1q} \\ \vdots & \ddots & \vdots \\ v^c_{q1} & \cdots & v^c_{qq} \end{pmatrix} + \begin{pmatrix} e_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e_{qq} \end{pmatrix}$$

ou encore :

$$V^s = \begin{pmatrix} v^c_{11} + e_{11} & \cdots & v^c_{1q} \\ \vdots & \ddots & \vdots \\ v^c_{q1} & \cdots & v^c_{qq} + e_{qq} \end{pmatrix} \tag{1.49}$$

La méthode des sommets apparaît comme une analyse inter-objets et intra-objets :

- c'est une analyse inter-objet puisqu'elle prend en compte dans l'analyse la variabilité entre les centres des hyper-rectangles (V^c).

- c'est une analyse intra-objet puisqu'elle prend en compte dans l'analyse la variabilité à l'intérieur des hyper-rectangles (facteur amplitude E).

1.5.2 La complexité des calculs

Dans la méthode des sommets, le calcul de m objets H_i décrits par q variables de type intervalle engendre une matrice de données X à $\sum_{i=1}^{n} 2^{q_i}$ lignes et q colonnes. La complexité étant en $O(2^q)$, il est clair que lorsque le nombre de variables descriptives augmente, les calculs deviennent coûteux.

Dans la méthode des centres, par contre, la complexité des calculs est en $O(q)$, c'est-à-dire du même ordre de complexité que dans le cas d'une Acp classique sur m objets décrits par des variables numériques.

Réduction de la complexité de *la méthode des sommets*

On a établi en 1.49 un résultat très important permettant de réduire la complexité de la méthode des sommets à l'ordre $O(q)$. En effet, au lieu de calculer la matrice de variance-covariance V^s à partir des descriptions des $\sum_{i=1}^{n} 2^{q_i}$ sommets, on déduit celle-ci en ajoutant au termes diagonaux de la matrice de variance-covariance V^c (calculées à partir des descriptions des centres des hyper-rectangles) les éléments diagonaux de la matrice diagonale E.
De plus les formules définies en 1.24 et 1.25 permettent d'obtenir les coordonnées factorielles de type intervalle sans avoir à calculer celles des 2^{q_i} sommets, ce qui réduit considérablement les temps de calcul.

1.6 La méthode des sommets du point de vue de la méthode Statis

Rappelons que dans la méthode des sommets on dispose de m tableaux $X_{H_1}, \ldots, X_{H_m}$. Chaque tableau X_{H_i} donne la description des 2^{p_i} sommets associés à l'objet H_i par q variables. On peut envisager d'utiliser la méthode Statis Duale pour analyser le tableau X des m groupes de sommets décrits par les mêmes variables :

$$X = \begin{pmatrix} X_{H_1} \\ \vdots \\ X_{H_m} \end{pmatrix}$$

On note V_i la matrice variance associée au tableau X_{H_i} :

$$V_i = Diag\left(\underline{p_{ij}}\overline{p_{ij}}(\overline{x_{ij}} - \underline{x_{ij}})^2/j = 1..q\right) \tag{1.50}$$

On définit un produit scalaire entre 2 matrices variances par la formule suivante :

$$\begin{aligned} < V_i, V_{i'} > &= \text{Trace}(V_i' V_{i'}) \\ &= \sum_{j=1}^{q}\sum_{l=1}^{q}(V_i)_{jl}(V_{i'})_{jl} = \sum_{j=1}^{q}(V_i)_{jj}(V_{i'})_{jj} \end{aligned}$$

puisque V_i et $V_{i'}$ sont diagonales, du fait de la symétrie des hyper-rectangles. On considère alors la matrice C $(m \times m)$ de terme général $c_{ii'} =< V_i, V_{i'} >$:

$$c_{ii'} = \sum_{j=1}^{q}(V_i)_{jj}(V_{i'})_{jj}$$

La diagonalisation de la matrice C permet de représenter l'ensemble des objets H_i à partir des vecteurs propres (normé à la racine carrée de la valeur propre) associés aux plus grandes valeurs propres. Soit $U = (u_1, \ldots, u_q)'$ le vecteur propre normé de C associé à la plus grande valeur propre. Alors le compromis V est défini par :

$$V = \sum_{i=1}^{m} u_i V_i$$

À partir de la diagonalisation de V on obtient une représentation des variables, ainsi qu'une représentation sur laquelle on peut projeter les $\sum_{i=1}^{m} 2^{p_i}$ lignes du tableau X. On peut noter que du fait de la structure parallèle des 2^{q_i} sommets propres à l'objet H_i, V_i est diagonale, donc V l'est, ce qui retire beaucoup d'intérêt à l'analyse des compromis et aux représentations qu'il permet.

1.7 La méthode des sommets du point de vue de l'analyse factorielle discriminante

Si l'objectif de l'analyse consiste, d'une part, à rechercher les axes factoriels qui séparent au mieux les hyper-rectangles, d'autre part, à identifier l'hyper-rectangle d'appartenance d'un nouvel individu, décrit par les variables descriptives, alors on peut appliquer une analyse factorielle discriminante (les classes étant les hyper-rectangles) à la matrice de donnée X définie en 1.3.

V_s la matrice de variance/covariance associée à X définit la variance totale, V_c (variance des centres) définit la variance *Inter-classes* et la variance *Intra-classes* est définie par la matrice diagonale V_E, égale à la moyenne des matrices de variance V_i diagonales, définies en 1.50.

La recherche des axes factoriels séparant au mieux les hyper-rectangles revient a diagonaliser la matrice non symétrique suivante :

$$V_s^{-1} V_c$$

ce qui revient après symétrisation à la diagonalisation de la matrice :

$$C V_s^{-1} C'$$

où $C' C = V_c$. Notant que cela correspond également à la diagonalisation de la matrice $V_E^{-1} V_c$, soit à la diagonalisation de la matrice symétrique $V_E^{\frac{-1}{2}} V_c V_E^{\frac{-1}{2}}$. $V_E^{\frac{-1}{2}}$ étant la matrice diagonale des inverses des racines carrées des éléments diagonaux de V_E.

1.8 Approche probabiliste

1.8.1 D'autres distributions à l'intérieur des intervalles

Dans l'approche probabiliste, on suppose que tout intervalle est le résumé d'un ensemble de valeurs qui se répartissent à l'intérieur de l'intervalle selon un loi connue. L'approche probabiliste revient à associer à chaque variable X_j et objet H_i une variable aléatoire x_{ij} de fonction de densité f_{ij} définie sur l'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$. On suppose ici l'indépendance entre les variables aléatoires x_{ij}. Soit le tableau de données suivant donnant la description des m objets $H_1, \ldots, H_m$ par q variables de type intervalle :

$$X_H = \begin{pmatrix} H_1 \\ \vdots \\ H_m \end{pmatrix} = \begin{array}{|c|c|c|} \hline [\underline{x_{11}}, \overline{x_{11}}] & \cdots & [\underline{x_{1q}}, \overline{x_{1q}}] \\ x_{11} \rightsquigarrow f_{11} & \cdots & x_{1q} \rightsquigarrow f_{1q} \\ \hline \vdots & \ddots & \vdots \\ \hline [\underline{x_{m1}}, \overline{x_{m1}}] & \cdots & [\underline{x_{mq}}, \overline{x_{mq}}] \\ x_{m1} \rightsquigarrow f_{m1} & \cdots & x_{mq} \rightsquigarrow f_{mq} \\ \hline \end{array} \tag{1.51}$$

Les figures suivantes montrent quelques exemples de lois de répartition symétriques à l'intérieur des intervalles.

Loi Uniforme U

Soit la variable aléatoire $x_{ij} \rightsquigarrow \mathcal{U}([\underline{x_{ij}}, \overline{x_{ij}}])$ qui suit une loi uniforme sur $[\underline{x_{ij}}, \overline{x_{ij}}]$. La fonction de densité f_{ij} et la fonction de répartition F_{ij} de x_{ij}, représentées dans la figure 1.12, ont pour expression :

$$f_{ij}(x) = \begin{cases} \frac{1}{(\overline{x_{ij}} - \underline{x_{ij}})} & \text{si x} \in [\underline{x_{ij}}, \overline{x_{ij}}] \\ 0 & \text{sinon} \end{cases}$$

$$F_{ij}(x) = \begin{cases} 0 & \text{si } x < \underline{x_{ij}} \\ \frac{(x - \underline{x_{ij}})}{(\overline{x_{ij}} - \underline{x_{ij}})} & \text{si x} \in [\underline{x_{ij}}, \overline{x_{ij}}] \\ 1 & \text{si } x > \overline{x_{ij}} \end{cases}$$

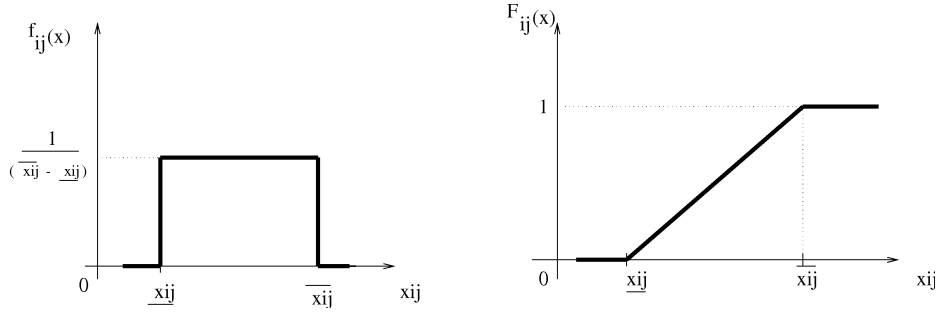


FIG. 1.12: $x_{ij} \rightsquigarrow \mathcal{U}([\underline{x_{ij}}, \overline{x_{ij}}])$

Loi Normale N[1]

Soit la variable aléatoire $x_{ij} \rightsquigarrow \mathcal{N}(\mu_{ij} = \frac{(\overline{x_{ij}}+\underline{x_{ij}})}{2}, \sigma_{ij} = \frac{(\overline{x_{ij}}-\underline{x_{ij}})}{6})$ qui suit une loi Normale de moyenne μ_{ij} et d'écart-type σ_{ij}. La fonction de densité f_{ij} a pour expression :

$$f_{ij}(x) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{\left(\frac{-1}{2}(\frac{(x-\mu_{ij})}{\sigma_{ij}})^2\right)}$$

Les fonctions de densité f_{ij} et de répartition F_{ij} de x_{ij} sont représentées dans la figure 1.13 suivante :

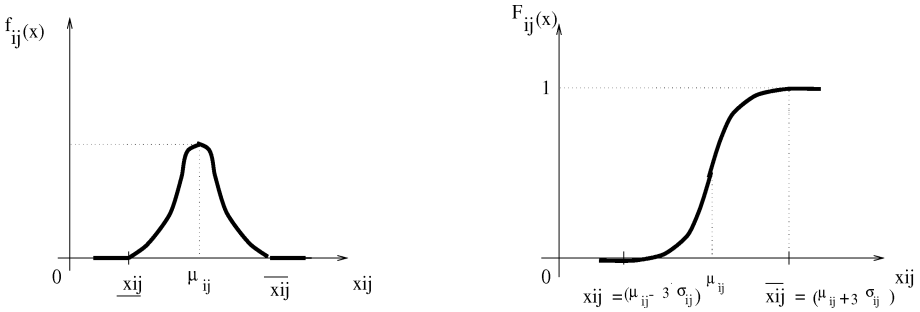


FIG. 1.13: $x_{ij} \rightsquigarrow \mathcal{N}(\mu = \frac{(\overline{x_{ij}}+\underline{x_{ij}})}{2}, \sigma = \frac{(\overline{x_{ij}}-\underline{x_{ij}})}{6})$

Loi Triangulaire T

Soit la variable aléatoire $x_{ij} \rightsquigarrow \mathcal{T}([\underline{x_{ij}}, \overline{x_{ij}}])$ qui suit une loi triangulaire sur $[\underline{x_{ij}}, \overline{x_{ij}}]$. La fonction de densité f_{ij} de x_{ij} a pour expression :

$$\begin{aligned} f_{ij}(x) &= \frac{4(x-\underline{x_{ij}})}{(\overline{x_{ij}}-\underline{x_{ij}})^2} \qquad \text{si } x \leq \frac{(\overline{x_{ij}}+\underline{x_{ij}})}{2} \\ &= \frac{4(\overline{x_{ij}}-x)}{(\overline{x_{ij}}-\underline{x_{ij}})^2} \qquad \text{si } x \geq \frac{(\overline{x_{ij}}+\underline{x_{ij}})}{2} \end{aligned}$$

Les fonctions de densité et de répartition sont représentées dans la figure 1.14 suivante :

1. En toute rigueur, il faudrait prendre une loi normale tronquée, nulle à l'extérieure de l'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$. Dans la mesure où la masse extérieure à cet intervalle est égale à $2.6^{10^{-3}}$ donc négligeable, nous ne tronquerons pas cette loi, afin de faciliter les calculs.

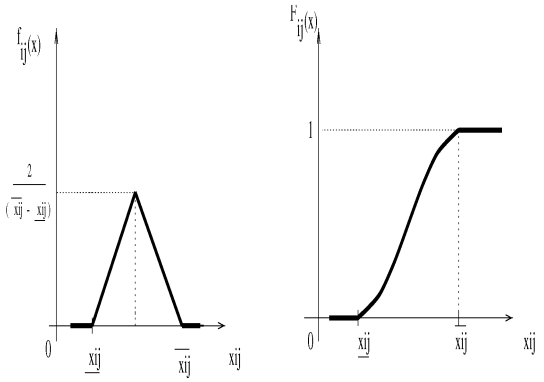


FIG. 1.14: $x_{ij} \rightsquigarrow \mathcal{T}([\underline{x_{ij}}, \overline{x_{ij}}])$

Loi Triangulaire inverse T'

Soit la variable aléatoire $x_{ij} \rightsquigarrow \mathcal{T}'([\underline{x_{ij}}, \overline{x_{ij}}])$ qui suit une loi triangulaire inverse sur $[\underline{x_{ij}}, \overline{x_{ij}}]$. La fonction de densité f_{ij} de x_{ij} a pour expression :

$$\begin{aligned} f_{ij}(x) &= \frac{2(\overline{x_{ij}} + \underline{x_{ij}} - 2x)}{(\overline{x_{ij}} - \underline{x_{ij}})^2} \qquad \text{si } x \leq \frac{(\overline{x_{ij}} + \underline{x_{ij}})}{2} \\ &= \frac{2(2x - \overline{x_{ij}} - \underline{x_{ij}})}{(\overline{x_{ij}} - \underline{x_{ij}})^2} \qquad \text{si } x \geq \frac{(\overline{x_{ij}} + \underline{x_{ij}})}{2} \end{aligned}$$

Les fonctions de densité et de répartition sont représentées dans la figure 1.15 suivante :

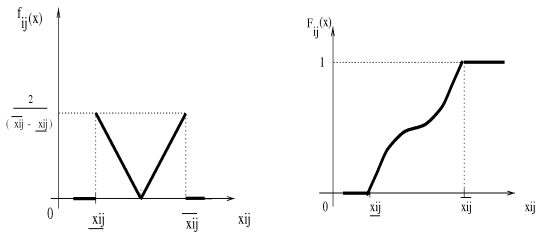


FIG. 1.15: $x_{ij} \rightsquigarrow \mathcal{T}'([\underline{x_{ij}}, \overline{x_{ij}}])$

1.8.2 La méthode des sommets et des centres du point de vue probabiliste

Du point de vue probabiliste, la méthode des sommets revient à répartir la masse d'un intervalle sur ces deux bornes $\underline{x_{ij}}$ et $\overline{x_{ij}}$ dans, respectivement, les proportions $\underline{p_{ij}}$ et $\overline{p_{ij}}$. La masse d'un intervalle, dans la méthode des centres, est supposée concentrée en son point de référence x_{ij}^o (masse de DIRAC en x_{ij}^o). Schématisons, dans le cas de la méthode des sommets et des centres, l'allure des fonctions de densité et de répartition de la distribution à l'intérieur des intervalles $[\underline{x_{ij}}, \overline{x_{ij}}]$.

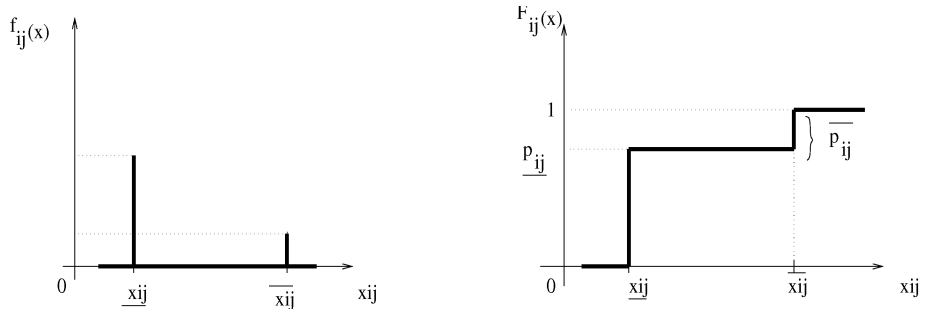


FIG. 1.16: *Distribution à l'intérieur d'un intervalle dans la méthode des sommets*

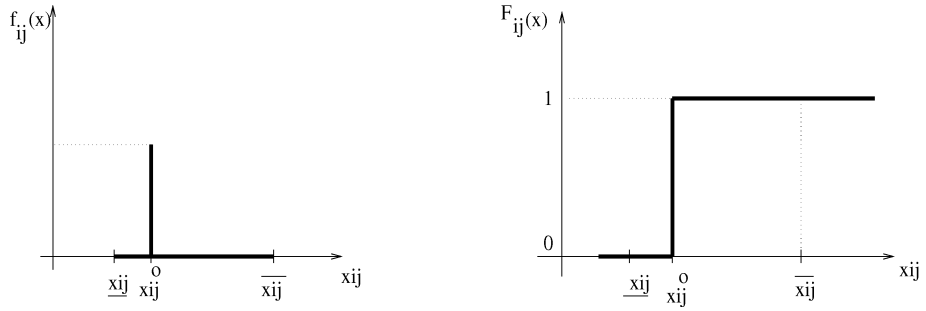


FIG. 1.17: *Distribution à l'intérieur d'un intervalle dans la méthode des centres*

1.8.3 Caractéristiques de tendance centrale et de dispersion des variables X_j

On note f_{X_j} la loi de la variable X_j définie comme suit :

$$f_{X_j} = \sum_{i=1}^{m} p_i f_{ij}$$

où p_i est le poids de l'objet H_i.

Les caractéristiques des variables X_j sont :

$$E(X_j) = \sum_{i=1}^{m} p_i \, E(x_{ij}) \tag{1.52}$$

$$\begin{aligned} V(X_j) &= E((X_j)^2) - E(X_j)^2 && (1.53)\\ &= \sum_{i=1}^{m} p_i E((x_{ij})^2) - \left(\sum_{i=1}^{m} p_i E(x_{ij}) \right)^2 && (1.54)\\ &= \sum_{i=1}^{m} p_i V(x_{ij}) + \sum_{i=1}^{m} p_i \left(E(x_{ij}) \right)^2 - \left(\sum_{i=1}^{m} p_i E(x_{ij}) \right)^2 && (1.55) \end{aligned}$$

En posant $E(x_{ij}) = x^o_{ij}$, alors $V(X_j) = \sum_{i=1}^{m} p_i V(x_{ij}) + v^c_{jj}$. v^c_{jj} étant la variance de la variable X_j dans la méthode des centres.

La loi jointe f_{X_j,X_l} des variables X_j, X_l est obtenue comme suit :

$$f_{X_j,X_l} = \sum_{i=1}^{m} p_i \, f^i_{x_{ij},x_{il}}$$

comme les variables aléatoires x_{ij} et x_{il} sont supposées indépendantes, on a :

$$f_{X_j,X_l} = \sum_{i=1}^{m} p_i \, f_{ij} \, f_{il}$$

Soit l'expression de la covariance entre deux variables X_j et X_l :

$$Cov(X_j, X_l) = E(X_j\, X_l) - E(X_j)E(X_l) \quad (1.56)$$
$$= \sum_{i=1}^{m} p_i(E(x_{ij})E(x_{il})) - \sum_{i=1}^{m} p_i E(x_{ij}) \sum_{i=1}^{m} p_i E(x_{il}) \quad (1.57)$$

En posant $E(x_{ij}) = x_{ij}^o$ et $E(x_{il}) = x_{il}^o$ alors $Cov(X_j, X_l) = v_{jl}^c = v_{jl}^s$.

1.8.4 Comparaison de la méthode des sommets et de la méthode des centres dans une approche probabiliste

1) Cas de lois uniformes

On suppose que chaque variable aléatoire x_{ij} suit une loi uniforme sur l'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$. Soit le tableau de données décrivant les m objets :

$$X_H = \begin{pmatrix} H_1 \\ \vdots \\ H_m \end{pmatrix} = \begin{array}{|c|c|c|} \hline x_{11} \rightsquigarrow \mathcal{U}([\underline{x_{11}}, \overline{x_{11}}]) & \cdots & x_{1q} \rightsquigarrow \mathcal{U}([\underline{x_{1q}}, \overline{x_{1q}}]) \\ \hline \vdots & \ddots & \vdots \\ \hline x_{m1} \rightsquigarrow \mathcal{U}([\underline{x_{m1}}, \overline{x_{m1}}]) & \cdots & x_{mq} \rightsquigarrow \mathcal{U}([\underline{x_{mq}}, \overline{x_{mq}}]) \\ \hline \end{array}$$

Comparons dans ce qui suit l'inertie prise en compte dans le cas des méthodes des sommets et des centres avec celle prise en compte dans le cas d'une distribution uniforme à l'intérieur des intervalles. Comparons pour cela les moyennes, variances et covariances des variables X_j dans le cas de ces trois méthodes.

Caractéristiques des variables x_{ij}

$$E(x_{ij}) = \int_{\underline{x_{ij}}}^{\overline{x_{ij}}} x \frac{1}{(\overline{x_{ij}} - \underline{x_{ij}})} = \frac{\overline{x_{ij}} + \underline{x_{ij}}}{2}$$
$$E((x_{ij})^2) = \int_{\underline{x_{ij}}}^{\overline{x_{ij}}} x^2 \frac{1}{\overline{x_{ij}} - \underline{x_{ij}}} = \frac{(\overline{x_{ij}})^2 + \overline{x_{ij}}\underline{x_{ij}} + (\underline{x_{ij}})^2}{3}$$
$$V(x_{ij}) = E((x_{ij})^2) - (E(x_{ij}))^2$$

$$= \frac{(\overline{x_{ij}} - \underline{x_{ij}})^2}{12}$$

Caractéristiques des variables X_j

Sachant que $E(x_{ij}) = x_{ij}^o = \frac{\overline{x_{ij}} + \underline{x_{ij}}}{2}$ alors $\underline{p_{ij}} = \overline{p_{ij}} = \frac{1}{2}$ et d'après les expressions établies en 1.52, 1.55 et 1.57 on a :

$$\begin{aligned} E(X_j) &= \overline{X^c} = \overline{X^s} \\ V(X_j) &= \sum_{i=1}^{m} p_i \frac{(\overline{x_{ij}} - \underline{x_{ij}})^2}{12} + v_{jj}^c \\ Cov(X_j, X_l) &= v_{jl}^s = v_{jl}^c \end{aligned}$$

2) Cas d'une loi normale N

On suppose que chaque variable aléatoire $x_{ij} \rightsquigarrow \mathcal{N}(\mu_{ij} = \frac{(\overline{x_{ij}} + \underline{x_{ij}})}{2}, \sigma_{ij} = \frac{(\overline{x_{ij}} - \underline{x_{ij}})}{6})$ suit une loi normale de moyenne μ_{ij} et d'écart-type σ_{ij}.

Caractéristiques des variables x_{ij}

$$\begin{aligned} E(x_{ij}) &= \mu_{ij} = \frac{\overline{x_{ij}} + \underline{x_{ij}}}{2} \\ V(x_{ij}) &= \sigma_{ij}^2 = \frac{(\overline{x_{ij}} - \underline{x_{ij}})^2}{36} \\ Cov(X_j, X_l) &= v_{jl}^s = v_{jl}^c \end{aligned}$$

Caractéristiques des variables X_j

$$\begin{aligned} E(X_j) &= \overline{X^s}_j = \overline{X^c}_j = \overline{X_j} \\ V(X_j) &= \sum_{i=1}^{m} p_i \frac{(\overline{x_{ij}} - \underline{x_{ij}})^2}{36} + v_{jj}^c \end{aligned}$$

3) Cas d'une loi triangulaire T

On suppose que chaque variable aléatoire $x_{ij} \rightsquigarrow \mathcal{T}([\underline{x_{ij}}\overline{x_{ij}}]$ suit une loi triangulaire sur l'intervalle $[\underline{x_{ij}}\overline{x_{ij}}]$.

Caractéristiques des variables x_{ij}

$$E(x_{ij}) = \frac{\overline{x_{ij}} + \underline{x_{ij}}}{2}$$

$$\text{où, } E(x_{ij}^2) = \frac{7}{24}\underline{x_{ij}}^2 + \frac{5}{12}\underline{x_{ij}}\,\overline{x_{ij}} + \frac{7}{24}\overline{x_{ij}}^2$$

$$V(x_{ij}) = \frac{(\overline{x_{ij}} - \underline{x_{ij}})^2}{24}$$

Caractéristiques des variables X_j

$$\begin{aligned} E(X_j) &= \overline{X_j} \\ V(X_j) &= v_{jj}^c + \sum_{i=1}^{m} p_i \frac{(\overline{x_{ij}} - \underline{x_{ij}})^2}{24} \\ Cov(X_j, X_l) &= v_{jl}^s = v_{jl}^c \end{aligned}$$

4) Cas d'une loi triangulaire inverse T

On suppose que chaque variable aléatoire $x_{ij} \rightsquigarrow \mathcal{T}'([\underline{x_{ij}}\overline{x_{ij}}]$ suit une loi triangulaire inverse sur l'intervalle $[\underline{x_{ij}}\overline{x_{ij}}]$.

Caractéristiques des variables x_{ij}

$$\begin{aligned} E(x_{ij}) &= \frac{\overline{x_{ij}} + \underline{x_{ij}}}{2} \\ E(x_{ij}^2) &= \frac{3}{8}\underline{x_{ij}}^2 + \frac{1}{4}\underline{x_{ij}}\,\overline{x_{ij}} + \frac{3}{8}\overline{x_{ij}}^2 \\ V(x_{ij}) &= \frac{1}{8}\left(\overline{x_{ij}} - \underline{x_{ij}}\right)^2 \end{aligned}$$

Caractéristiques des variables X_j

$$\begin{aligned} E(X_j) &= \overline{X_j} \\ V(X_j) &= \sum_{i=1}^{m} p_i \frac{\left(\overline{x_{ij}} - \underline{x_{ij}}\right)^2}{8} + v_{jj}^c \\ Cov(X_j, X_l) &= v_{jl}^s = v_{jl}^c \end{aligned}$$

On récapitule les résultats précédents dans un tableau comparatif. On y compare les expressions des moyennes, des variances et des covariances des variables X_j, dans le cas de certaines lois usuelles, dans le cas de la méthode des centres et celle des sommets. On suppose ce qui ne restreint pas la généralité, que les variables X_j sont centrées ($\forall\ j\ \ \overline{X_j} = o$) :

	Moyenne	Variance	Covariance
Méthode des sommets	$\overline{X_j^s} = \overline{X_j} = 0$	$v_{jj}^s = \sum_{i=1}^m p_i \overline{p_{ij}} \underline{p_{ij}} \, (\overline{x_{ij}} - \underline{x_{ij}})^2$ $+v_{jj}^c$	$v_{jl}^s = v_{jl}^c$
Méthode des centres	$\overline{X_j^c} = \sum_{i=1}^m p_i(\overline{p_{ij}}\,\overline{x_{ij}} + \underline{p_{ij}}\,\underline{x_{ij}})$ $= \overline{X_j} = 0$	$v_{jj}^c = \sum_{i=1}^m p_i \left(\overline{p_{ij}}\,\overline{x_{ij}} + \underline{p_{ij}}\,\underline{x_{ij}}\right)^2$	$Cov(X_j, X_l) = \sum_{i=1}^m p_i(\overline{p_{ij}}\,\overline{x_{ij}} + \underline{p_{ij}}\,\underline{x_{ij}}) \cdot$ $(\overline{p_{il}}\,\overline{x_{il}} + \underline{p_{il}}\,\underline{x_{il}})$
Loi Uniforme	$E(X_j) = \overline{X_j} = 0$	$V(X_j) = \sum_{i=1}^m p_i \frac{1}{12} \, (\overline{x_{ij}} - \underline{x_{ij}})^2$ $+v_{jj}^c$ où $\underline{p_{ij}} = \overline{p_{ij}} = \frac{1}{2}$	$Cov(X_j, X_l) = v_{jl}^c$
Loi Normale	$E(X_j) = \overline{X_j} = 0$	$V(X_j) = \sum_{i=1}^m p_i \frac{1}{36} \, (\overline{x_{ij}} - \underline{x_{ij}})^2$ $+v_{jj}^c$ où $\underline{p_{ij}} = \overline{p_{ij}} = \frac{1}{2}$	$Cov(X_j, X_l) = v_{jl}^c$
Loi $\mathcal{T}$	$E(X_j) = \overline{X_j} = 0$	$V(X_j) = v_{jj}^c + \frac{\sum_{i=1}^m p_i(\overline{x_{ij}} - \underline{x_{ij}})^2}{24}$ où $\underline{p_{ij}} = \overline{p_{ij}} = \frac{1}{2}$	$Cov(X_j, X_l) = v_{jl}^c$
Loi $\mathcal{T}'$	$E(X_j) = \overline{X_j} = 0$	$V(X_j) = v_{jj}^c + \frac{\sum_{i=1}^m p_i(\overline{x_{ij}} - \underline{x_{ij}})^2}{8}$ où $\underline{p_{ij}} = \overline{p_{ij}} = \frac{1}{2}$	$Cov(X_j, X_l) = v_{jl}^c$

Discussion

On constate, d'une part, que les lois considérées laissent les moyennes et les covariances constantes. D'autre part, les variances se décomposent en la

somme de deux effets :

- **un effet fixe** correpondant à la variance V^c, due à la variance des centres des hyper-rectangles (variance inter-objets),
- **un effet variable** estimant, en fonction de la distribution, la variation intrinsèque à chaque hyper-rectangle (variance intra-objets).

Les variances peuvent être classées par ordre croissant comme suit :

$$V^{c*} < V^{\mathcal{T}'} < \quad V^{\mathcal{T}} < V^{\mathcal{N}} < V^{s*}$$

où V^{s*} et V^{c*} sont les variances V^s et V^c associées à une pondération équirépartie : $(\overline{p_{ij}} = \underline{p_{ij}} = \frac{1}{2})$.

1.9 Approche concurrente

En s'inspirant des travaux de CALAHAN [Calahan72] et de NAGARAJ [Nagaraj93], NAGABUSHAN [Nagabhushan97] propose une extension de l'analyse en composantes principales à des données intervalles. Nous introduisons d'abord, les travaux de CALAHAN et de NAGARAJ puis l'algorithme proposé par NAGABUSHAN étendant l'ACP aux données intervalles.

Les travaux de CALAHAN [Calahan72] porte sur l'analyse de la tolérance des circuits électriques. Un circuit électrique est caractérisé par un ensemble de paramètres d'entrée $p_1, \ldots, p_r$ et un ensemble de paramètres de sortie $z_1, \ldots z_n$. Chaque paramètre de sortie z_i est une fonction des paramètres d'entrées $p_1, \ldots, p_r$. L'analyse de la tolérance d'un circuit consiste à observer le comportement extrême du circuit en faisant varier les paramètres p_j.

Soit $[\underline{p_j}, \overline{p_j}]$ l'intervalle de variation du paramètre p_j, $p_j^o \in [\underline{p_j}, \overline{p_j}]$ une valeur dite nominale autour de laquelle on observe la variation et z_i^o la sortie du circuit correspondante aux entrées $(p_1^o, \ldots, p_r^o)$. Le problème de l'analyse de la tolérance revient à déterminer les valeurs extrêmes $\underline{z_i}$ et $\overline{z_i}$ de chaque sortie z_i sachant que chaque paramètre p_j varie dans l'intervalle $[\underline{p_j}, \overline{p_j}]$. Pour cela, CALAHAN proposa d'utiliser les dérivées partielles des fonctions z_i par rapport aux paramètres p_j (développements en séries de Taylor) pour déterminer la variation maximale autour de z_i^o, soit :

$$\Delta z_i \quad = \quad \left(\sum_{i=1}^{r} \Delta_{p_j} \left(\frac{\delta z_i}{\delta p_j} \right) \quad / p_j = p_j^o \right) \tag{1.58}$$

où Δz_i est la variation de z_i autour de la valeur nominale z_i^o et $\Delta p_j = (p_j - p_j^o)$.

Comme la dérivée $\dfrac{\delta z_i}{\delta p_j}$ peut être une expression positive ou négative, alors la variation maximale positive notée Δz_i^+ est obtenue en choisissant :

$$\begin{aligned} p_j = \underline{p_j} \qquad & \text{si} \frac{\delta z_i}{\delta p_j} < 0 \ \text{ et} \\ p_j = \overline{p_j} \qquad & \text{si} \frac{\delta z_i}{\delta p_j} > 0 \end{aligned}$$

De manière similaire, la variation maximale négative notée Δz_i^- est obtenue en choisissant :

$$\begin{aligned} p_j = \underline{p_j} \qquad & \text{si} \frac{\delta z_i}{\delta p_j} > 0 \quad \text{et} \\ p_j = \overline{p_j} \qquad & \text{si} \frac{\delta z_i}{\delta p_j} < 0 \end{aligned}$$

Les valeurs extrêmes de z_i sont alors :

$$\begin{aligned} \underline{z_i} \quad &= \quad z_i^o + \Delta z_i^+ \\ \overline{z_i} \quad &= \quad z_i^o + \Delta z_i^- \end{aligned}$$

1.9.1 La réduction de dimension sur des données intervalles

En s'inspirant des travaux de Calahan, Nagaraj [Nagaraj93] proposa un algorithme générique de réduction de dimension permettant de décrire, dans un espace de dimension faible, un ensemble d'objets décrits par des variables de type intervalle. Cet algorithme fut appliqué par Nagabhushan [Nagabhushan97] au cas particulier d'une analyse en composantes principales. On présente dans ce qui suit l'algorithme de réduction de dimension appliqué

à des données intervalles, puis son application dans le cadre d'une Acp.

Position du problème

Soit m objets décrits par q variables numériques $X_1, \ldots, X_q$, et une méthode de réduction de dimension définie par r fonctions $(f_1, \ldots, f_r)$ $(r < q)$. Ces fonctions permettent de décrire l'ensemble des objets par un nombre réduit de nouvelles variables $Y_1, \ldots, Y_r$ définies comme suit :

$$\forall\, j = 1..r \qquad\qquad Y_j \;=\; f_j(X_1, \ldots, X_q)$$

Nagaraj propose d'étendre de telles méthodes de réduction de dimension à des objets décrits par des variables de type intervalle. Soit m objets décrits par q variables de type intervalle $X_1, \ldots, X_q$:

$$\begin{pmatrix} 1 \\ \vdots \\ m \end{pmatrix} \;=\; \begin{pmatrix} [\underline{x_{11}}, \overline{x_{11}}] & \cdots & [\underline{x_{1q}}, \overline{x_{1q}}] \\ \vdots & \ddots & \vdots \\ [\underline{x_{m1}}, \overline{x_{m1}}] & \cdots & [\underline{x_{mq}}, \overline{x_{mq}}] \end{pmatrix} \tag{1.59}$$

L'objectif est de décrire l'ensemble des objets dans un espace de dimension plus faible défini par de nouvelles variables $Y_1, \ldots, Y_r$ également de type intervalle :

$$\begin{pmatrix} 1 \\ \vdots \\ m \end{pmatrix} \;=\; \begin{pmatrix} [\underline{y_{11}}, \overline{y_{11}}] & \cdots & [\underline{y_{1r}}, \overline{y_{1r}}] \\ \vdots & \ddots & \vdots \\ [\underline{y_{m1}}, \overline{y_{m1}}] & \cdots & [\underline{y_{mr}}, \overline{y_{mr}}] \end{pmatrix} \tag{1.60}$$

Principe de l'algorithme proposé

Le principe de cet algorithme se fonde sur la dérivation des fonctions de réduction de dimension $f_1, \ldots, f_r$ pour déduire la variation maximale $[\underline{y_{ij}}, \overline{y_{ij}}]$ prise par la nouvelle variable Y_j pour l'objet i. Deux cas sont envisagés. Dans le cas où les fonctions f_i admettent des expressions analytiques connues, les variations des nouvelles variables sont obtenues par dérivation des fonctions f_i en utilisant la formule développée en 1.58. Dans le cas où les fonctions f_i n'admettent pas d'expressions connues, NAGARAJ propose une procédure estimant les dérivées des fonctions f_i par rapport aux variables Y_j. Ces estimations sont ensuite utilisées dans la formule 1.58 pour approximer les variations des nouvelles variables.

A) Fonctions f_i connues

1- On considère la matrice X^o suivante où chaque objet est décrit par son point de référence x_i^o :

$$X^o = \begin{pmatrix} x_1^o \\ \vdots \\ x_m^o \end{pmatrix} = \begin{pmatrix} x_{11}^o & \cdots & x_{1q}^o \\ \vdots & \ddots & \vdots \\ x_{m1}^o & \cdots & x_{mq}^o \end{pmatrix} \tag{1.61}$$

où $x_{ij}^o \in [\underline{x_{ij}}, \overline{x_{ij}}]$ est une valeur de référence autour de laquelle est observée la variation. En général, on considère $x_{ij}^o = \dfrac{\overline{x_{ij}} + \underline{x_{ij}}}{2}$.

2- On applique la méthode de réduction de dimension à la matrice des descriptions X^o. Soit $Y_1^o, \ldots, Y_r^o$ les nouvelles variables donnant la description dans un espace réduit des points x_i^o :

$$\begin{pmatrix} x_1^o \\ \vdots \\ x_m^o \end{pmatrix} = \begin{pmatrix} y_{11}^o & \cdots & y_{1r}^o \\ \vdots & \ddots & \vdots \\ y_{m1}^o & \cdots & y_{mr}^o \end{pmatrix} \tag{1.62}$$

3- Pour chaque objet et pour chaque variable Y_j on détermine la variation maximale positive ΔY_{ij}^+ et la variation maximale négative ΔY_{ij}^- sur Y_j pour l'objet i, par rapport au point nominal x_i^o :

$$\Delta Y_{ij}^{+} = \frac{\delta(f_j)}{\delta X_1}\Delta X_{i1} + \ldots, + \frac{\delta(f_j)}{\delta X_q}\Delta X_{iq}$$

où $\frac{\delta(f_j)}{\delta X_i}$ est la dérivée de la fonction f_j par rapport à la variable X_i au point x_i^o et $\Delta X_{ij} = (x_{ij} - x_{ij}^o)$. Comme la dérivée $\dfrac{\delta f_j}{\delta X_i}$ peut être une expression positive ou négative, alors la variation maximale positive ΔY_{ij}^{+} est obtenue en choisissant $x_{ij} = \underline{x_{ij}}$ si $\dfrac{\delta f_j}{\delta X_i} < 0$ et $x_{ij} = \overline{x_{ij}}$ si $\dfrac{\delta f_j}{\delta X_i} > 0$.

De manière similaire, la variation négative maximale ΔY_{ij}^{-} est obtenue en choisissant $x_{ij} = \underline{x_{ij}}$ si $\dfrac{\delta f_j}{\delta X_i} > 0$ et $x_{ij} = \overline{x_{ij}}$ si $\dfrac{\delta f_j}{\delta X_i} < 0$.

4- Les valeurs extrêmes de y_{ij} de la variable X_j pour un objet i sont :

$$\begin{aligned} \underline{y_{ij}} &= y_{ij}^o + \Delta Y_{ij}^{+} \\ \overline{y_{ij}} &= y_{ij}^o + \Delta Y_{ij}^{-} \end{aligned}$$

$\dfrac{\delta(f_j)}{\delta X_i}$ représente la proportion de variation de la variable Y_j engendrée par une variation de la variable X_i.

B) Fonctions f_i non connues

Dans le cas où les fonctions f_i ne sont pas bien définies, on estime les dérivées partielles en procédant, pour chaque variable X_j, à deux perturbations : une dans le sens positif et l'autre dans le sens négatif. Pour chaque perturbation de la variable X_j, on étudie la perturbation engendrée sur les nouvelles variables en sortie.

a) On perturbe dans le sens positif la variable X_1 de α (pour les calculs $\alpha = 10\%$). Soit $X^{(1)+}$ la nouvelle matrice obtenue :

$$X^{(1)+} = \begin{pmatrix} 1 \\ \vdots \\ m \end{pmatrix} = \begin{pmatrix} x_{11}^o + \alpha\Delta x_{11} & \cdots & x_{1q}^o \\ \vdots & \ddots & \vdots \\ x_{m1}^o + \alpha\Delta x_{m1} & \cdots & x_{mq}^o \end{pmatrix} \tag{1.63}$$

où $\Delta x_{i1} = \overline{x_{i1}} - x^o_{i1}$.

b) On applique la méthode de réduction de dimension à la nouvelle matrice obtenue après perturbation. Soit la nouvelle description des objets dans l'espace défini par les nouvelles variables $Y_1^{(1)+}, \ldots, Y_r^{(1)+}$:

$$\begin{pmatrix} 1 \\ \vdots \\ m \end{pmatrix} = \begin{pmatrix} y_{11}^{(1)+} & \cdots & y_{1r}^{(1)+} \\ \vdots & \ddots & \vdots \\ y_{m1}^{(1)+} & \cdots & y_{mr}^{(1)+} \end{pmatrix} \tag{1.64}$$

L'estimation $S^{f_j}_{X_1^+}$ de la dérivée partielle $\dfrac{\delta(f_j)}{\delta X_1}$ de la fonction f_j par rapport à la variable perturbée positivement X_1 est définie comme suit :

$$S^{f_j}_{X_1^+} = \frac{\sum_{i=1}^m (y_{ij}^{1+} - y_{ij}^o)}{\sum_{i=1}^m (x_{i1}^{(1)+} - x_{i1}^o)}$$

avec $x_{i1}^{(1)+} = x_{i1}^o + \alpha \Delta x_{i1}$ et y_{ij}^o, rappelons-le, est la valeur prise par la nouvelle variable Y_j^o (à la suite de la réduction de dimension) pour l'objet i.

c) De manière similaire, on perturbe X_1^o dans le sens négatif comme suit :

$$X^{(1)-} = \begin{pmatrix} 1 \\ \vdots \\ m \end{pmatrix} = \begin{pmatrix} x_{11}^o - \alpha \Delta x_{11} & \cdots & x_{1q}^o \\ \vdots & \ddots & \vdots \\ x_{m1}^o - \alpha \Delta x_{m1} & \cdots & x_{mq}^o \end{pmatrix} \tag{1.65}$$

où $\Delta x_{i1} = (x_{i1}^o - \underline{x_{i1}})$.

d) On applique la méthode de réduction de dimension à la nouvelle matrice. Soit la nouvelle description des objets dans l'espace défini par les variables $Y_1^{(1)-}, \ldots, Y_r^{(1)-}$:

$$\begin{pmatrix} 1 \\ \vdots \\ m \end{pmatrix} = \begin{pmatrix} y_{11}^{(1)-} & \cdots & y_{1r}^{(1)-} \\ \vdots & \ddots & \vdots \\ y_{m1}^{(1)-} & \cdots & y_{mr}^{(1)-} \end{pmatrix} \tag{1.66}$$

L'estimation $S_{X_1^-}^{f_j}$ de la dérivée partielle $\frac{\delta(f_j)}{\delta X_1}$ par rapport à la variable X_1 perturbée négativement est définie comme suit :

$$S_{X_1^-}^{f_j} = \frac{\sum_{i=1}^{m}(y_{ij}^{(1)-} - y_{ij}^{o})}{\sum_{i=1}^{m}(x_{i1}^{(1)-} - x_{i1}^{o})}$$

avec $x_{i1}^{(1)-} = x_{i1}^{o} - \alpha\Delta x_{i1}$.

e) On réitère les étapes a,b,c,d successivement pour les variables $X_2, \ldots, X_q$.

f) On calcule pour chaque objet i et pour chaque variable Y_j la variation maximale positive ΔY_{ij}^{+} et la variation maximale négative ΔY_{ij}^{-} sur Y_j, par rapport au point nominal x_i^o :

$$\Delta Y_{ij}^{+} = \sum_{j=1}^{q} S_{X_j^+}^{f_j}(x_{ij} - x_{ij}^{o})$$

avec

$$\begin{aligned} x_{ij} &= \underline{x_{ij}} \text{ si } S_{X_j^+}^{f_j} < 0 \text{ au point } x_i^o \text{ et} \\ x_{ij} &= \overline{x_{ij}} \text{ si } S_{X_j^+}^{f_j} > 0 \end{aligned}$$

De manière similaire, ΔY_{ij}^{-} est défini comme suit :

$$\Delta Y_{ij}^{-} = \sum_{j=1}^{q} S_{X_j^-}^{f_j}(x_{ij} - x_{ij}^{o})$$

avec

$$\begin{aligned} x_{ij} &= \overline{x_{ij}} \text{ si } S_{X_j^-}^{f_j} < 0 \text{ au point } x_i^o \text{ et} \\ x_{ij} &= \underline{x_{ij}} \text{ si } S_{X_j^-}^{f_j} > 0 \end{aligned}$$

Les valeurs extrêmes prises par la variable Y_j pour l'objet i sont définies comme suit :

$$\begin{aligned} \underline{y_{ij}} &= y_{ij}^o + \Delta Y_{ij}^+ \\ \overline{y_{ij}} &= y_{ij}^o + \Delta Y_{ij}^- \end{aligned}$$

1.9.2 Application de l'algorithme dans le cas d'une ACP

NAGABUSHAN [Nagabhushan97] a appliqué l'algorithme de réduction de dimension étendu aux intervalles dans le cas où la méthode de réduction de dimension est l'analyse en composantes principales. Il a introduit quelques modifications dans le calcul de l'estimation des dérivées partielles.

Dans le cas de l'analyse en composantes principales, les variables $Y_1, \ldots, Y_r$ définissent les r premières composantes principales. Les fonctions f_i, dans le cas de variables X_i numériques centrées, sont alors :

$$Y_j = f_j(X_1, \ldots, X_q) = \sum_{l=1}^{q} x_{il} u_{lj}$$

Or les composantes des vecteurs propres u_{lj} ne sont connues qu'après la diagonalisation de la matrice de variance-covariance. On se situe alors dans le cas où les fonctions f_i sont inconnues. On procède alors à l'estimation des dérivées partielles pour ensuite approximer les intervalles de variation de chaque objet sur les r premiers axes factoriels.

L'algorithme de l'ACP sur des données intervalles

1- Soit X^o la matrice $(m \times q)$ donnant la description des m objets par leur point de référence (ou par leur centre) :

$$X^o = \begin{pmatrix} x_1^o \\ \vdots \\ x_m^o \end{pmatrix} = \begin{pmatrix} x_{11}^o & \cdots & x_{1q}^o \\ \vdots & \ddots & \vdots \\ x_{m1}^o & \cdots & x_{mq}^o \end{pmatrix} \tag{1.67}$$

2- On applique l'analyse en composantes principales au nuage des centres défini par la matrice X^o. La description des centres dans l'espace des r premières composantes principales $Y_1^o, \ldots, Y_r^o$ est :

$$\begin{pmatrix} x_1^o \\ \vdots \\ x_m^o \end{pmatrix} = \begin{pmatrix} y_{11}^o & \cdots & y_{1r}^o \\ \vdots & \ddots & \vdots \\ y_{m1}^o & \cdots & y_{mr}^o \end{pmatrix} \tag{1.68}$$

3 - On perturbe positivement la variable X_k. Soit $X^{(k+)}$ la matrice obtenue après perturbation.

4 - On applique l'Acp aux données perturbées définies par $X^{(k+)}$.

5 - On calcule pour chaque objet i et pour chaque fonction f_j l'estimation $S^{f_j}_{X_k^+}(i)$ de la dérivée partielle de la fonction f_j par rapport à la variable X_k perturbée positivement, comme suit :

$$S^{f_j}_{X_k^+}(i) = \frac{(y_{ij}^{(k+)} - y_{ij}^o)}{(x_{ik}^{(k+)} - x_{ik}^o)} = s_{ij}^{(k+)}$$

où $s_{ij}^{(k+)}$ est la proportion de la variation de la $j^{ème}$ coordonnée factorielle y_{ij} de l'objet i à la suite d'une perturbation positive de la variable X^k.

6 - De manière similaire, on perturbe la variable X_k dans le sens négatif. Soit $X^{(k-)}$ la matrice obtenue après perturbation.

7 - On applique l'Acp aux données perturbées négativement définies par $X^{(k-)}$.

8 - On calcule pour chaque objet i l'estimation $S^{f_j}_{X_k^-}(i)$ de la dérivée partielle de la fonction f_j par rapport à la variable X_k perturbée négativement ; soit :

$$S^{f_j}_{X_k^-}(i) = \frac{(y_{ij}^{(k-)} - y_{ij}^o)}{(x_{ik}^{(k-)} - x_{ik}^o)} = s_{ij}^{(k-)}$$

9 - On réitère les étapes de 1 à 8 successivement pour les variables $X_2, \ldots, X_q$.

10 - On calcule pour chaque objet i et pour chaque variable Y_j la variation maximale positive ΔY_{ij}^{+} et la variation maximale négative ΔY_{ij}^{-} sur Y_j, par rapport au point nominal x_i^o comme suit :

$$\Delta Y_{ij}^{+} = \sum_{k=1}^{q} S_{X_k^+}^{f_j}(i)(x_{ij} - x_{ij}^o)$$

avec

$$\begin{aligned} x_{ij} &= \underline{x_{ij}} \text{ si } S_{X_k^+}^{f_j}(i) < 0 \text{ au point } x_i^o \text{ et} \\ x_{ij} &= \overline{x_{ij}} \text{ si } S_{X_k^+}^{f_j}(i) > 0 \end{aligned}$$

De manière similaire, ΔY_{ij}^{-} est défini comme suit :

$$\Delta Y_{ij}^{-} = \sum_{k=1}^{q} S_{X_k^-}^{f_j}(i)(x_{ij} - x_{ij}^o)$$

avec

$$\begin{aligned} x_{ij} &= \overline{x_{ij}} \text{ si } S_{X_k^-}^{f_j}(i) < 0 \text{ au point } x_i^o \text{ et} \\ x_{ij} &= \underline{x_{ij}} \text{ si } S_{X_k^-}^{f_j}(i) > 0 \end{aligned}$$

Les valeurs extrêmes prises par la $j^{ème}$ composante principale Y_j pour l'objet i sont définies comme suit :

$$\begin{aligned} \underline{y_{ij}} &= y_{ij}^o + \Delta Y_{ij}^{+} \\ \overline{y_{ij}} &= y_{ij}^o + \Delta Y_{ij}^{-} \end{aligned}$$

1.9.3 Discussion

Résumons le principe de cet algorithme. La première partie consiste à effectuer une ACP sur le nuage des centres des intervalles X^o ; chaque objet est ensuite visualisé par son centre dans l'espace factoriel. La seconde partie consiste à estimer la variation maximale autour de chaque centre projeté dans l'espace factoriel.

Pour ce faire, on calcule pour chaque variable X_k deux matrices $X^{(k+)}$, $X^{(k-)}$, issues de la perturbation respectivement positive et de la perturbation négative de la variable X^k. On applique sur chacune des matrices perturbées une ACP. On calcule ensuite deux matrices $S^{(k+)}(s_{ij}^{k+})$ et $S^{(k-)}(s_{ij}^{k-})$ à m lignes et r colonnes, où le terme général, s_{ij}^{k+} estime la proportion de la variation de la $j^{ème}$ coordonnée factorielle y_{ij} de l'objet i à la suite d'une perturbation positive de X^k. Finalement, on déduit à partir des $2q$ matrices $S^{(k+)}$ et $S^{(k-)}$ $(k = 1..q)$, c'est-à-dire à partir des $2mrq$ ratios, la variation maximale autour des centres des objets dans l'espace factoriel.

Remarquons que la première partie de cet algorithme correspond à l'approche de la méthode des centres, où les axes factoriels sont définis à partir des centres et ne tiennent pas compte de la variation exprimée par les intervalles. Contrairement à la méthode des centres, l'estimation de la variation autour des centres implique dans cet algorithme des calculs assez lourds. En effet, pour estimer la variation de tous les objets, on doit appliquer l'ACP $2q$ fois et estimer $2mrq$ ratios. Tout en sachant de plus que les variations finales estimées ne sont correctes, de par les propriétés des dérivées partielles, que lorsque les amplitudes des intervalles de départ sont très faibles. Or il paraît évident que, pour restituer la variation autour des centres dans l'espace factoriel, il suffit de projeter en tant qu'éléments supplémentaires l'ensemble des sommets des hyper-rectangles puis de tracer la "couverture rectangulaire" reliant les sommets d'un même objet. C'est ce qui se fait dans la méthode des centres.

De plus, afin d'optimiser les calculs on propose les formules 1.45 et 1.46 qui permettent de fournir les limites de variation, dans l'espace factoriel, autour de chaque centre sans avoir à calculer les coordonnées factorielles de l'ensemble des sommets. Ce qui réduit considérablement les temps de calcul.

1.10 Une méthode symbolique de génération de classes d'hyper-rectangles

Afin d'introduire cette méthode, rappelons le formalisme de base de l'analyse des données symboliques :

E : l'ensemble des individus ,

$\mathcal{D}$: l'ensemble des descriptions ,

X : une application de E dans $\mathcal{D}$, qui associe à chaque individu de E sa description. Soit :

$$\begin{array}{rll} X: & E & \longrightarrow \mathcal{D} \\ & e & \longrightarrow X(e) \end{array}$$

$\mathcal{R}$: une relation entre deux descriptions. Si $d \in \mathcal{D}$ et $d' \in \mathcal{D}$ sont en relation par $\mathcal{R}$, on note $d\mathcal{R}d'$. Plus généralement, on peut considérer que R est une relation "floue" autrement dit que d et d' peuvent être plus au moins en relation. On note $[d'\mathcal{R}d]$ le degré de relation liant d à d'. En analyse des données symbolique, $\mathcal{R}$ joue le rôle d'opérateur de comparaison ou d'appariement entre descriptions comme $=, \leq, \geq, \subseteq$, etc. Le résultat de cette comparaison est à valeur dans L. L'ensemble L est défini soit par $\{0, 1\}$ soit par l'intervalle $[0, 1]$.

$\mathcal{A}$: une application de E dans L qui associe à un individu $e \in E$ un degré d'adéquation entre sa description et une description $d \in \mathcal{D}$. $\mathcal{A}$ est aussi dite fonction de reconnaissance. Soit :

$$\begin{array}{rll} \mathcal{A}: & E & \longrightarrow L \\ & e & \longrightarrow [X(e) \ \ \mathcal{R} \ \ d] \end{array}$$

Un objet symbolique est défini par la donnée du triplet $(\mathcal{A}, \mathcal{R}, d)$. Selon la nature de L on peut définir deux types d'objets symboliques :

- *Les objets symboliques booléens* c'est le cas où l'ensemble L est égal à {0,1}.

- *Les objets symboliques modaux* c'est le cas où l'ensemble L est égal à $[0,1]$.

Une assertion booléenne est un cas particulier d'objet symbolique, où chaque individu e de E possède p descriptions par rapport aux p variables $X_1,\ldots,X_p$; l'espace des descriptions $\mathcal{D}$ est donné par $D_1,\ldots,D_p$. $\mathcal{R}$ est défini à l'aide d'un vecteur de relations $(\mathcal{R}_1,\ldots,\mathcal{R}_p)$ par $d'\,\mathcal{R}\,d = \wedge_{i=1..p}[d'_i\,\mathcal{R}\,d_i]$ où le $\wedge$ est la conjonction logique, autrement dit, qui prend la valeur 1 si $[d'_i\,\mathcal{R}\,d_i] = 1 \quad \forall\ i = 1..p$ et 0 sinon. La fonction de reconnaissance est une application $\mathcal{A}$ de E dans $\{0,1\}$:

$$\mathcal{A}(e) \;=\; \wedge_{i=1}^{p}[X_i(e)\ \mathcal{R}_i\ d_i]$$

où d_i, $1 \leq i \leq p$, est une partie de D_i. Remarquons que la donnée de cette égalité ($\wedge$) permet de définir sans ambiguïté l'objet symbolique $s = (\mathcal{A},\mathcal{R},d)$.

L'union symbolique $\oplus$ aussi appelé jonction [Ichino et al.94] est un opérateur permettant de généraliser deux descriptions.

$$\begin{array}{rll} \oplus : & E \times E & \longrightarrow \mathcal{D} \\ & (e_i,e_j) & \longrightarrow X(e_i) \oplus X(e_j) = (X_1(e_i) \cup X_1(e_j),\ldots,X_1(e_i) \cup X_p(e_j)) \end{array}$$

où $\cup$ est un opérateur d'union (qui est en général une t-conorme) associé au descripteur D_k.

Si X_k est qualitative, alors $\cup$ est l'opérateur d'union ensembliste.

Si X_k est quantitative, soit $X_k(e_i) = d_{ik} = [\underline{x_{ik}},\overline{x_{ik}}]$ et $X_k(e_j) = d_{jk} = [\underline{x_{jk}},\overline{x_{jk}}]$, alors $\cup$ est défini comme suit :

$$\begin{array}{rcl} d_{ik} \cup d_{jk} & = & \{d_{ik},d_{jk}\}, \quad \text{si } d_{ik} \text{ et } d_{jk} \text{ sont des intervalles disjoints} \\ & = & [min(\underline{x_{ik}},\underline{x_{jk}}),max(\underline{x_{ik}},\underline{x_{jk}})] \quad \text{sinon} \end{array}$$

Génération de classes d'hyper-rectangles dans l'espace factoriel

On note X l'application de E dans $\mathcal{D}$ qui associe à chaque hyper-rectangle $H_i \in E$ sa description, soit :

$$\begin{array}{rlll} X: & E & \longrightarrow (X_1^I \times \ldots \times X_p^I) \\ & H_i & \longrightarrow X(H_i) = (X_1(H_i), \ldots, X_p(H_i)) \end{array}$$

On associe à l'hyper-rectangle H_i l'objet symbolique S_i défini par le triplet $(\mathcal{A}, \subseteq, X(H_i))$. Soit $\mathcal{F}$ étant l'espace factoriel des q premières composantes principales $(PC_1, \ldots, PC_q)$ $(q \leq p)$ et PC_k^I étant l'ensemble des intervalles associés à la $k^{ème}$ composante principale PC_k. On note PC l'application de E dans $\mathcal{F}$ qui associe à chaque hyper-rectangle H_i sa description factorielle, soit :

$$\begin{array}{rlll} PC: & E & \longrightarrow (PC_1^I \times \ldots \times PC_q^I) \\ & H_i & \longrightarrow PC(H_i) = (PC_1(H_i), \ldots, PC_q(H_i)) \end{array}$$

Cette méthode simple de génération des classes d'hyper-rectangles est fondée sur l'union symbolique. Soit $\mathcal{P}(E)$ l'ensemble des parties de E et $H \in \mathcal{P}(E)$ un ensemble d'hyper-rectangles à généraliser. La description de la classe H dans l'espace factoriel des q premières composantes principales est :

$$\begin{array}{rlll} PC: & \mathcal{P}(E) & \longrightarrow PC_1^I \times \ldots \times PC_q^I \\ & H & \longrightarrow PC(H) = (\oplus_{H_i \in H} PC_1(H_i), \ldots, \oplus_{H_i \in H} PC_q(H_i)) \\ & & \quad = (PC_1(H), \ldots, PC_q(H)) \end{array}$$

L'objet symbolique associé à la classe d'hyper-rectangles H est défini par le triplet $(\mathcal{A}, \subseteq, PC(H))$.

Extension de l'objet symbolique H

Dans le cas booléen, l'extension de la classe d'hyper-rectangles H est définie comme suit :

$$Ext(H) = \{H_k \in E \,/\, \mathcal{A}(H_k) = \wedge_{i=1}^{q} [PC_i(H_k) \subseteq PC_i(H)] = 1\}$$

Dans le cas modal, l'extension de H est alors,

$$Ext(H) = \{H_k \in E \ / \ \mathcal{A}(H_k) = \wedge_{i=1}^{q}[PC_i(H_k) \subseteq_f PC_i(H)] \geq \alpha\}$$

où $\subseteq_f$ est une fonction d'inclusion floue [Grabisch et al.95] et $0 \leq \alpha \leq 1$.

Chapitre 2

Codage flou des variables intervalles en vue d'une ACM

Résumé du chapitre

Utilisée pour le traitement des questionnaires et l'exploitation des enquêtes, l'analyse des correspondances multiples constitue l'essentiel du succès de l'analyse des correspondances auprès des praticiens. À défaut de méthodes pouvant analyser des données complexes, les questionnaires sont souvent conçus de telles façon que les sujets sont contraints de répondre par une seule valeur pour chaque question.

Nous nous intéressons dans ce travail à des types de données complexes naturellement rencontrés dans les données d'enquête : les intervalles.

On propose trois techniques de codage des variables de type intervalle : le codage croisé, le codage par sommets et le codage sans décomposition. Les deux premières techniques se fondent sur la décomposition des variables intervalles en variables numériques.

La dernière technique se fonde sur l'extension d'outils de codage des variables numériques à des données intervalles. Pour cela, on s'intéresse, dans une première partie, à l'extension des techniques classiques de découpage à des variables de type intervalle. Dans une deuxième partie, on s'intéresse aux fonctions d'appartenance qui mesurent le degré d'appartenance d'un intervalle à une classe d'intervalles. Après l'étude des mesures de distance et de dissimilarité entre les intervalles et la définition des propriétés que doit vérifier une fonction d'appartenance associée à une classe d'intervalles, on propose une fonction d'appartenance fondée sur la distance de MOORE.

2.1 Introduction

Dans ce chapitre, notre intérêt porte sur une autre méthode tout aussi importante en analyse factorielle : l'analyse des correspondances multiples (ACM).

L'ACM classique permet d'étudier la structure d'une population d'objets décrits par des variables qualitatives et/ou quantitatives. L'ACM est bien adaptée au traitement des questionnaires, à l'exploitation des enquêtes lorsque toutes les variables sont qualitatives ou que l'on a transformé les variables quantitatives en variables qualitatives ordinales.

Deux problématiques sont toujours plus ou moins explicitement présentes dans les objectifs d'une analyse des correspondances multiples.

D'une part, il y a la problématique d'une ACP dont le but est de chercher une typologie des individus ainsi qu'une typologie des variables. D'autre part, il y a la problématique d'une analyse des correspondances dont le but est de chercher une typologie des modalités en étudiant la liaison entre toutes les variables. Les trois typologies individus, variables et modalités sont ensuite mises en correspondance.

2.1.1 Le codage des données en ACM

La première étape en ACM consiste à transformer, à l'aide de codages appropriés, les variables quantitatives et qualitatives en des variables binaires ou floues. Une variable X est définie comme une fonction de l'ensemble Ω des objets à valeurs dans le domaine O_X :

$$\begin{array}{rcl} X : \Omega & \rightarrow & O_X \\ \omega & \rightarrow & X(\omega) \end{array}$$

où $X(\omega)$ est la valeur de la variable X prise par l'objet ω. Le codage de la variable X par la variable X' de domaine $O_{X'}$ définit une fonction de Ω à valeurs dans $O_{X'}$:

$$\begin{array}{rcl} X' : \Omega & \rightarrow & O_{X'} \\ \omega & \rightarrow & X'(\omega) \end{array}$$

X' est aussi la composée de la fonction X et d'une application c définie sur O_X à valeurs dans $O_{X'}$:

$$\begin{array}{cccccc} & & X & & c & \\ X' : & \Omega & \rightarrow & O_X & \rightarrow & O_{X'} \\ & \omega & \rightarrow & X(\omega) & \rightarrow & c(X(\omega)) = coX(\omega) = X'(\omega) \end{array}$$

Le codage d'une variable X en une variable X' en ACM peut être décomposé en deux étapes successives :

- le découpage du domaine O(X) de la variable X en classes ;
- le calcul du degré d'appartenance des objets aux classes définies précédemment.

Codage disjonctif d'une variable quantitative

1. Découpage du domaine de la variable

On distingue trois principales techniques de découpage du domaine d'une variable quantitative en classes :

- découpage en classes d'amplitudes égales ;
- découpage en classes d'effectifs égaux ;
- découpage en classes minimisant l'inertie.

a) Découpage en classes d'amplitudes égales

Le découpage du domaine O_X d'une variable X en k classes $C_1, \ldots, C_k$ d'amplitudes égales est essentiellement utilisé dans le cas d'une distribution uniformément répartie sur O_X, les classes obtenues sont de densités relativement équivalentes.

Dans le cas d'une distribution non uniforme, le découpage en classes d'amplitudes égales risque de construire des classes de densité très faibles ou vides. Dans ce cas sont utilisés d'autres types de découpage, tel que le le découpage en classes d'effectifs égaux, par histogramme, ou encore par minimisation de l'inertie.

b) Découpage en classes d'effectifs égaux

La première étape consiste à discrétiser le domaine O_X en p intervalles $I_1, \ldots, I_p$ avec $I_j = [a_{j-1}, a_j[$. On note nj le nombre de valeurs de la variable X appartenant à l'intervalle I_j. Soit F la fonction de répartition définie aux bornes a_j avec $F(a_j)$ la fréquence cumulée au point a_j.

On note $C_1, \ldots, C_k$ avec $C_j = [\underline{c_j}, \overline{c_j}[$, les classes d'effectifs égaux à définir. Connaissant la fréquence cumulée $\frac{j}{k}$ au point $\overline{c_j}$, il s'agit de déterminer la valeur $\overline{c_j}$ pour $j = 1, .., m-1$. L'intervalle de variation $I_r = [a_{r-1}, a_r[$ contenant la valeur $\overline{c_j}$ vérifie :

$$F(a_{r-1}) \leq \frac{j}{k} \leq F(a_r)$$

La valeur $\overline{c_j}$ est obtenue par interpolation comme suit :

$$\overline{c_j} = \frac{(\frac{j}{k} - F(a_{r-1}))(a_r - a_{r-1})}{F(a_r) - F(a_{r-1})} + a_{r-1}$$

Le critère de découpage à effectifs égaux ne tient compte, lors de la construction des classes, que des effectifs dans les classes et non de la proximité des valeurs.

c) Découpage par minimisation de l'inertie

Soit $\mathcal{P}_k(O_X)$ l'ensemble des partitions à k classes défini sur O_X. Soit W un critère de mesure de la **qualité d'une partition**, il définit une application sur $\mathcal{P}_k(O_X)$ à valeurs dans $\Re^+$:

$$\begin{array}{rccc} W : & \mathcal{P}_k(O_X) & \rightarrow & \Re^+ \\ & p_i & \rightarrow & W(p_i) \end{array}$$

On utilise souvent l'inertie d'une partition comme mesure de qualité. On définit $\mathcal{I}(p)$ l'inertie d'une partition p comme la somme des inerties des classes qui la composent. Soit $p = (C_1, \ldots, C_k)$ une partition de O_X en k classes, l'inertie de cette partition est ainsi calculée :

$$\mathcal{I}(p) = \sum_{i=1}^{k} card(C_i)\mathcal{I}(C_i) = \sum_{i=1}^{k} card(C_i) \frac{(\sum_{\omega \in C_i} (X(\omega) - \overline{X_i})^2)}{\sum_{i=1}^{k} card(C_i)}$$

où

$$\overline{X_i} = \frac{\sum_{\omega \in C_i} X(\omega)}{card(C_i)}$$

Une partition $p\star \in \mathcal{P}_k(O_X)$ est dite **optimale** au sens d'un critère W si et seulement si :

$$\forall p \in P_k(O_X) \;\; W(p\star) \leq W(p)$$

Pour définir une telle partition on peut utiliser l'algorithme de FISHER [Fisher58] qui fournit une partition optimale sur un ensemble V doté d'une structure d'ordre total et d'une mesure de ressemblance ou de dissemblance entre ses éléments.

2. Fonctions d'appartenance aux classes

Soit $C_1, \ldots, C_k$ les k classes issues du découpage du domaine $O(X)$ de la variable X. On associe à chaque classe C_j une fonction d'appartenance φ_j, dite aussi fonction caractéristique à valeur dans $\{0, 1\}$, définie comme suit :

$$\begin{array}{rrcl} \varphi_j : & O_X & \rightarrow & \{0,1\} \\ & X(\omega) & \rightarrow & \varphi_j(X(\omega)) = 1 \;\text{ si }\; X(\omega) \in C_j \\ & & & \quad = 0 \;\text{ sinon} \end{array}$$

Le codage disjonctif c de la variable X en k modalités est :

$$\begin{array}{rrcl} c : & O_X & \rightarrow & \{0,1\}^k \\ & X(\omega) & \rightarrow & (\varphi_1(X(\omega)), \ldots, \varphi_k(X(\omega))) \end{array}$$

Les fonctions φ_j vérifient la propriété suivante :

$$\forall\, \omega \;\in\; \Omega \qquad \sum_{i=1}^{k} \varphi_i(X(\omega)) \;\; = 1 \tag{2.1}$$

Codage disjonctif d'une variable qualitative

Le découpage du domaine d'une variable qualitative en classes (modalités) peut être défini par un algorithme de classification ou par un expert. Soit $C_1, \ldots, C_k$ les k classes. À chaque classe C_j est associée une fonction caractéristique φ_j, indiquant la présence ou l'absence de la modalité prise par l'objet ω dans la classe C_j.

Remarquons que dans le cas où toutes les classes sont des singletons, chaque classe C_j identifie la modalité j de la variable, et la fonction φ_j est la fonction caractéristique de la modalités j de la variable X. Les fonctions φ_j vérifient également la propriété 2.1.

2.1.2 L'Acm sous codage disjonctif

Soit T le tableau issu du codage disjonctif des variables qualitatives et quantitatives. Il donne la description des m objets par p variables binaires $X_1, \ldots, X_p$ à respectivement $q_1, \ldots, q_p$ modalités :

$$
T = \begin{array}{c} \\ 1 \\ \vdots \\ m \end{array}
\begin{array}{|c|c|c|c|c|c|c|}
\multicolumn{2}{c}{X_1} & & & \multicolumn{2}{c}{X_p} & \\
\hline
m_1^1 & \cdots & m_{q_1}^1 & \cdots & m_1^p & \cdots & m_{q_p}^p \\
\hline
t_{11}^1 & \cdots & t_{11}^{q_1} & \cdots & t_{1p}^1 & \cdots & t_{1p}^{q_p} \\
\hline
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\hline
t_{m1}^1 & \cdots & t_{m1}^{q_1} & \cdots & t_{mp}^1 & \cdots & t_{mp}^{q_p} \\
\hline
\end{array}
$$

où $t_{ij}^k \in \{0, 1\}$ peut être interprété comme la probabilité pour l'individu i de prendre la modalité k de la variable X_j.

$$
\begin{aligned}
\sum_{k=1}^{q_j} t_{ij}^k &= 1 \qquad \forall i \in \{1..m\} \quad \forall j \in \{1..p\} \\
q_j &: \text{ est le nombre de modalités de la variable } X_j \,; \\
q = \sum_{j=1}^{p} q_j &: \text{ le nombre des modalités toute variable confondue.}
\end{aligned}
$$

Les marges en lignes du tableau disjonctif complet sont constantes et égales au nombre de variables p :

$$t_{i.} = \sum_{j=1}^{p} \sum_{k=1}^{q_k} t_{ij}^k \quad = \quad p$$

Les marges en colonnes correspondant au nombre d'objets caractériés par la modalité k de la variable j :

$$t_{.j}^k = \sum_{i=1}^{m} t_{ij}^k$$

Inertie prise en compte

Les objets sont tous dotés d'un poids identique égal à $p_i = \frac{1}{m}$. La $k^{ème}$ modalité m_j^k de la variable j est de poids $p_j^k = \frac{t_{.j}^k}{p\,m}$. La distance entre la modalité m_j^k et le centre de gravité du nuage G, dont toutes les m coordonnées valent $\frac{1}{m}$, s'écrit :

$$d^2(j,G) \quad = \quad m \sum_{i=1}^{m} \left(\frac{t_{ij}^k}{t_{.j}^k} - \frac{1}{m} \right)^2 = \frac{m}{t_{.j}^k} - 1$$

La distance d'une modalité au centre de gravité est d'autant plus grande que l'effectif est plus faible.

Inertie d'une modalité

L'inertie $I(m_j^k)$ de la modalité m_j^k vaut :

$$I(m_j^k) \quad = \quad p_j^k \; d^2(j,G) = \frac{1}{p}(1 - \frac{t_j^k}{m})$$

La contribution d'une modalité à l'inertie totale est d'autant plus grande que l'effectif (le poids) dans cette modalité est plus faible.

Inertie d'une variable

L'inertie $I(X_j)$ de la variable X_j, vaut :

$$I(X_j) = \sum_{k=1}^{q_j} I(m_j^k) = \frac{1}{p}(q_j - 1)$$

La part d'inertie due à une variable est proportionnelle au nombre de ses modalités.

Inertie totale

On en déduit l'inertie totale I :

$$I = \sum_{j=1}^{p} I(X_j) = \frac{q}{p} - 1$$

Ainsi, l'inertie totale dépend uniquement du nombre de variables et de modalités et non des liaisons entre les variables.

2.1.3 Le codage flou

Le codage disjonctif est un codage en classes (en modalités) bien séparées fournissant des résultats clairs et facilement interprétables. Il est bien adapté aux analyses non-linéaires puisqu'il permet de décrire des rapports non linéaires entre les variables.

Cependant on distingue plusieurs limites au codage disjonctif. D'une part, il y a une perte d'information quand une valeur d'une variable est codée par son appartenance ou pas à un intervalle donné. D'autre part, ce codage est irréversible, dans le sens où on ne peut pas reconstituer les données d'origines à partir des données codées en disjonctif. La plus importante limite du codage disjonctif réside dans sa discontinuité : il y a création d'une distance artificielle entre deux valeurs très proches mais appartenant à deux intervalles contigus. Pour remédier à cela on a eu recours au codage flou.

Le codage flou a suscité de nombreux travaux. Le codage fonctionnel par morceaux remonte aux travaux de Bordet [Bordet73] en 1973. Guitonneau et Roux [Guitonneau et al.77] présentent une application d'un codage trapézoïdal et comparent le résultat avec celui d'une analyse en composantes principales pondérée, Ghermani, Roux et Roux [Ghermani et al.77] proposent et appliquent des codages flous sous des considérations empiriques. Lafaye de Micheaux [Lafayedemicheaux78] est le premier à avoir établi le lien entre les codages linéaires par morceaux et les travaux théoriques sur l'approximation des analyses continues non-linéaires. Martin [Martin80] établit un parallèle entre certaines formes de codages flous et les méthodes d'estimation de la densité (lissage d'histogramme). Le Foll [Lefoll79] étend le codage flou à des graphes non-hiérarchiques, aux mesures de proximités et au codage polynomial.

En analyse des correspondances multiples, les travaux sur les codages flous remonte à Gallego [Gallego82]. On note également ceux de Cazes [Cazes90] et de Rijckevorsel [Rijckevorsel88] [Rijckevorsel87].

Pour pallier à la discontinuité introduite par le codage disjonctif en Acm Gallego propose un codage progressif d'une modalité à l'autre conservant des résultats clairs et une interprétation lisible en Acm. Il propose le codage semi-linéaire, où les modalités sont caractérisées par des fonctions d'appartenance valant 1 aux points de référence et décroissant linéairement d'un côté et de l'autre.

2.1.4 L'Acm sous codage flou

Soit T le tableau issu du codage flou et/ou disjonctif des variables quantitatives et/ou qualitatives. Il donne la description des m objets par p variables $X_1, \ldots, X_p$ à respectivement $q_1, \ldots, q_p$ modalités :

$$
T = \begin{array}{c} \\ 1 \\ \vdots \\ m \end{array}
\begin{array}{c}
\begin{array}{cccccccc} & X_1 & & & & & X_p & \end{array} \\
\begin{array}{|c|c|c|c|c|c|c|}
\hline
m_1^1 & \cdots & m_{q_1}^1 & \cdots & m_1^p & \cdots & m_{q_p}^p \\ \hline
t_{11}^1 & \cdots & t_{11}^{q_1} & \cdots & t_{1p}^1 & \cdots & t_{1p}^{q_p} \\ \hline
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \hline
t_{m1}^1 & \cdots & t_{m1}^{q_1} & \cdots & t_{mp}^1 & \cdots & t_{mp}^{q_p} \\ \hline
\end{array}
\end{array}
$$

où $0 \leq t_{ij}^k \leq 1$ peut être interprété comme la probabilité pour l'individu i de prendre la modalité k de la variable X_j. On note,

$$\begin{aligned} \sum_{k=1}^{q_j} t_{ij}^k &= 1 \qquad \forall i \in \{1..m\} \quad \forall j \in \{1..p\} \\ q_j &: \text{ est le nombre de modalités de la variable } X_j\,; \\ q = \sum_{j=1}^{p} q_j &: \text{ le nombre des modalités toute variable confondue.} \end{aligned}$$

Dans le cas particulier où toutes les valeurs t_{ij}^k prises par la variable X_j sont égales à 0 ou à 1, la variable X_j est dite binaire. Dans le cas contraire elle est dite floue.

Inertie prise en compte

L'inertie totale, les contributions des variables et des modalités à cette inertie demeurent toujours plus faibles dans le cas d'un codage flou que celles obtenues dans le cas d'un codage disjonctif. Considérons l'expression suivante de l'inertie d'une modalité :

$$\begin{aligned} I(m_j^k) &= p_j^k \; d^2(j,G) = p_j^k \; m \sum_{i=1}^{m} \left(\frac{t_{ij}^k}{t_{.j}^k} - \frac{1}{m} \right)^2 \\ &= m \left(\frac{\sum_{i=1}^{m} (t_{ij}^k)^2}{(t_{.j}^k)^2} - \frac{1}{m} \right) \end{aligned}$$

Dans le cas disjonctif, comme t_{ij}^k est égal à 0 ou à 1, alors $(t_{ij}^k)^2 = t_{ij}^k$; l'expression précédente se simplifie et on obtient $I(m_j^k) = \dfrac{m}{t_{.j}^k} - 1$.

Dans le cas d'un codage flou, comme $0 \leq t_{ij}^k \leq 1$ alors $(t_{ij}^k)^2 \leq t_{ij}^k$; par conséquent, l'inertie d'une modalité est toujours inférieure ou égale à celle obtenue dans le cas disjonctif :

$$I(m_j^k) = m \frac{\sum_{i=1}^{m} (t_{ij}^k)^2}{(t_{.j}^k)^2} - 1 \leq m \frac{\sum_{i=1}^{m} t_{ij}^k}{(t_{.j}^k)^2} - 1 = \frac{m}{t_{.j}^k} - 1$$

Ainsi, l'inertie totale (égale à la somme des inerties des modalités toute variable confondue) obtenue dans le cas d'un codage flou est toujours inférieure ou égale à celle obtenue dans le cas d'un codage disjonctif.

2.1.5 Positionnement du travail

Dans la même problématique que celle du chapitre 1, l'objectif du présent travail est d'étendre l'ACM à des tableaux de données impliquant des variables hétérogènes de type qualitative, quantitative et intervalle. Pour ce faire, on propose de nouvelles techniques de codage flou de variables de type intervalle. On aborde le codage flou d'une variable intervalle en deux parties : d'abord, le découpage de la variable en classes, ensuite la définition des fonctions d'appartenance associées aux classes.

Outre les problèmes classiques du découpage d'une variable continue en classes (nombre de classes, choix des bornes des classes, construction de classes non creuses, etc.), on soulève dans le cas des variables intervalles deux questions fondamentales. La première porte sur le choix de l'interprétation à associer aux classes : une classe peut représenter, comme dans le cas numérique, une région de valeurs (basses, moyennes ou hautes), elle peut désigner un ensemble d'intervalles satisfaisant certaines propriétés (classe d'intervalles de faibles amplitudes et situés dans des régions de faibles valeurs, des intervalles de grandes amplitudes situés dans des régions de faibles ou moyennes valeurs ...), etc. La deuxième question porte sur la définition d'un découpage conservant l'interprétation choisie.

Après le découpage de la variable intervalle en classes, il s'agit de définir les fonctions d'appartenance permettant de mesurer le degré d'appartenance des intervalles observés aux classes. La notion d'appartenance est très liée à la sémantique de la classe. Par exemple, dans le cas où une classe désigne une région de valeurs, l'appartenance d'un intervalle à cette classe (elle même un intervalle) revient à une mesure de distance ou de similarité entre intervalles. Il existe un grand nombre de mesures de distance (similarité) définies entre les intervalles, mais il est nécessaire que celles-ci fournissent des résultats pertinents et interprétables en ACM. Ces réflexions autour de la sémantique d'une classe, le découpage et la définition des fonctions d'appartenance, nous ont mené à proposer trois techniques de codage flou d'une variable intervalle.

Le principe des deux premières techniques, dites **Codage croisé** et **Codage par sommets**, est la transformation de la variable intervalle en variables numériques. Les variables numériques obtenues sont ensuite codées en utilisant les techniques classiques de codage. Finalement, on reconstitue le codage de chaque variable intervalle à partir du codage des variables numériques qui la composent.

Le **codage croisé** décompose chaque variable intervalle en deux variables numériques. Le codage fl), des codages flous des deux variables numériques qui la composent. Le codage croisé fournit un codage de tendance centrale tenant compte de l'information de variation exprimée par les données intervalles. La variation intrinsèque à chaque objet n'est pas visualisée, dans les plans factoriels, elle est prise en compte lors du codage de l'objet.

Le **codage par sommets** utilise une décomposition verticale portant sur les objets. Chaque ligne donnant la description d'un objet par au moins une variable de type intervalle est décomposée en un ensemble de lignes donnant la description des sommets associés à l'objet en question. Le codage de l'objet est déduit à partir des codages de ses sommets. Le codage par sommets fournit une information sur la variation intrinsèque à chaque objet. Il permet, contrairement au codage croisé, la visualisation de la variation dans les plans factoriels.

Dans la troisième technique de codage dit **sans décomposition**, plutôt que de transformer les variables intervalles en variables numériques, on propose d'étendre les outils classiques de codage à des variables de type intervalle. Pour le découpage en classe, on propose une généralisation de la notion d'histogramme et de fonction de répartition à des distributions d'intervalles.
Après la définition des propriétés que doit vérifier une fonction d'appartenance (évolution de l'appartenance en fonction de l'amplitude, de la position, etc.) on propose une fonction basée sur la distance de Moore.

2.2 Codage croisé d'une variable de type intervalle

Le codage croisé d'une variable intervalle consiste à la décomposer en deux variables numériques (*décomposition horizontale*) définies par l'ensemble des bornes inférieures et l'ensemble des bornes supérieures des intervalles observés. Les deux variables numériques sont ensuite codées à l'aide d'une technique de codage appropriée : le domaine de chaque variable numérique est découpé en classes et à chaque classe est associée une fonction d'appartenance.

Le découpage du domaine d'une variable intervalle en classes est obtenu par le croisement des classes issues du découpage des domaines de chacune des variables numériques. De manière similaire, les fonctions d'appartenance associées aux classes de la variable intervalle sont définies par le croisement des fonctions d'appartenance associées aux classes issues du découpage des variables numériques.

Nous détaillons dans ce qui suit chacune de ces étapes, nous présentons ensuite quelques cas particuliers de codage croisé : *le codage binaire croisé* et *le codage semi-linéaire croisé*. Enfin, nous illustrons la technique du codage croisé à l'aide d'un exemple.

2.2.1 Décomposition de la variable intervalle

Soit X une variable de type intervalle observée sur m objets et $[\underline{x_i}, \overline{x_i}]$ l'intervalle observé pour l'objet S_i. On note $[x_{min}, x_{max}]$ le domaine de variation de la variable X où les bornes x_{min} et x_{max} sont définies comme suit :

$$\begin{aligned} x_{min} &= \min_{(i=1..m)} \underline{x_i} \\ x_{max} &= \max_{(i=1..m)} \overline{x_i} \end{aligned}$$

La première étape du codage consiste à décomposer la variable X en deux variables X_{min}, X_{max} définies comme suit :

$$\begin{aligned} X_{min} &= \{ \underline{x_i} \ / \ i = 1..m \} \\ X_{max} &= \{ \overline{x_i} \ / \ i = 1..m \} \end{aligned}$$

Soit la description des objets après la décomposition de la variable X :

	X	
	X_{min}	X_{max}
S_1	$\underline{x_1}$	$\overline{x_1}$
$\vdots$	$\vdots$	$\vdots$
S_m	$\underline{x_m}$	$\overline{x_m}$

TAB. 2.1: *Description des objets S_i après décomposition*

Une représentation possible des intervalles pris par une variable X de type intervalle de visualiser chaque intervalle $[\underline{x_i}, \overline{x_i}]$, dans le plan défini par les deux variables X_{min} et X_{max}, par un point de coordonnées $\underline{x_i}$ et $\overline{x_i}$. Sachant que $\underline{x} \leq \overline{x}$, alors tout intervalle $[\underline{x}, \overline{x}]$ est représenté par un point de coordonnée $(\underline{x}, \overline{x})$ situé dans le demi-plan supérieur (P_1).

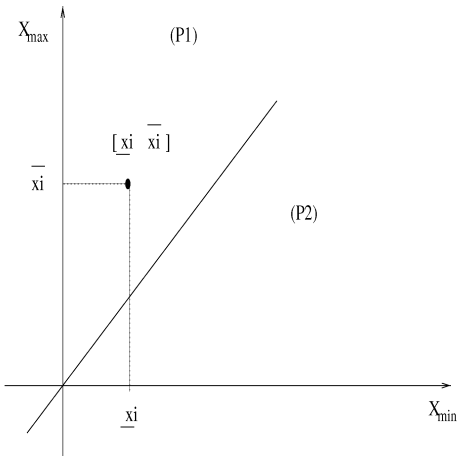


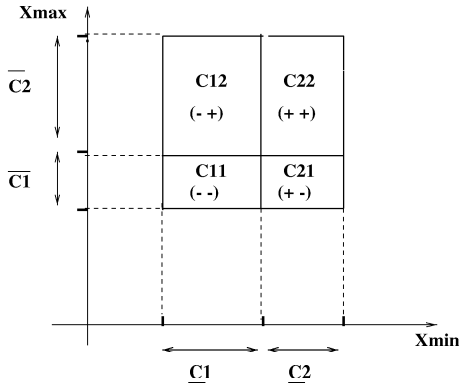
FIG. 2.1: *Représentation ponctuelle d'un intervalle*

2.2.2 Découpage d'une variable intervalle

On note $\{\underline{C_1}, \ldots, \underline{C_{k_{min}}}\}$ et $\{\overline{C_1}, \ldots, \overline{C_{k_{max}}}\}$, respectivement, les k_{min} et les k_{max} classes issues du découpage classique des variables X_{min} et X_{max}. On définit la classe C_{rl} de la variable X issue du croisement des classes $\underline{C_r}$ et $\overline{C_l}$ comme l'ensemble des intervalles $[\underline{x_i}, \overline{x_i}]$ tel que :

$$\underline{x_i} \in \underline{C_r} \quad \text{et} \quad \overline{x_i} \in \overline{C_l}$$

Un croisement entre une classe $\underline{C_r}$ et une classe $\overline{C_l}$ est dit valide s'il existe au moins une valeur de la classe $\underline{C_r}$ inférieure à au moins une valeur de la classe $\overline{C_l}$. Dans le cas où $k_{min} = 2$ et $k_{max} = 2$, le découpage de la variable X est illustré par la figure suivante :



Les classes $\underline{C_1} = [\underline{a_1}, \overline{a_1}]$ et $\underline{C_2} = [\underline{a_2}, \overline{a_2}]$ (resp. $\overline{C_1} = [\underline{b_1}, \overline{b_1}]$ et $\overline{C_2} = [\underline{b_2}, \overline{b_2}]$) découpent le domaine de la variable X_{min}, X_{max} en deux régions, l'une de faibles valeurs et l'autre de fortes valeurs. Les classes (modalités) C_{rl} découpent le domaine de la variable X non seulement selon le positionnement des intervalles (valeurs faibles, fortes ...) mais également selon leur amplitudes. Dans le cas d'un découpage en quatre classes valides les interprétations associées aux modalités de la variable X sont :

Modalités	Domaine de variation des intervalles	Amplitude des intervalles
C_{11}	$[\underline{a_1}, \overline{b_1}]$	$[max(0, (\underline{b_1} - \overline{a_1})), max(0, (\overline{b_1} - \underline{a_1}))]$
C_{21}	$[\underline{a_2}, \overline{b_1}]$	$[max(0, (\underline{b_1} - \overline{a_2})), max(0, (\overline{b_1} - \underline{a_2}))]$
C_{12}	$[\underline{a_1}, \overline{b_2}]$	$[max(0, (\underline{b_2} - \overline{a_1})), max(0, (\overline{b_2} - \underline{a_1}))]$
C_{22}	$[\underline{a_2}, \overline{b_2}]$	$[max(0, (\underline{b_2} - \overline{a_2})), max(0, (\overline{b_2} - \underline{a_2}))]$

Ainsi, si un objet est situé dans le plan factoriel proche, par exemple, de la modalité C_{22} alors, d'une part, l'intervalle pris par cet objet pour la variable

X varie dans la région des fortes valeurs $[\underline{a_2}, \overline{b_2}]$, d'autre part, son amplitude est faible et varie dans $[max(0, (\underline{b_2} - \overline{a_2})), max(0, (\overline{b_2} - \underline{a_2}))]$. En revanche, un objet situé proche de la classe C_{12} présente une grande variation sur tout le domaine $[\underline{a_1}, \overline{b_2}]$ et caractérisé, par conséquent, par une large amplitude variant dans $[max(0, (\underline{b_2} - \overline{a_1})), max(0, (\overline{b_2} - \underline{a_1}))]$.

Un autre type de représentation consiste à projeter l'ensemble des intervalles pris par la variable X sur un axe ; le découpage des variables X_{min} et X_{max} en 2 classes chacune est représenté comme suit :

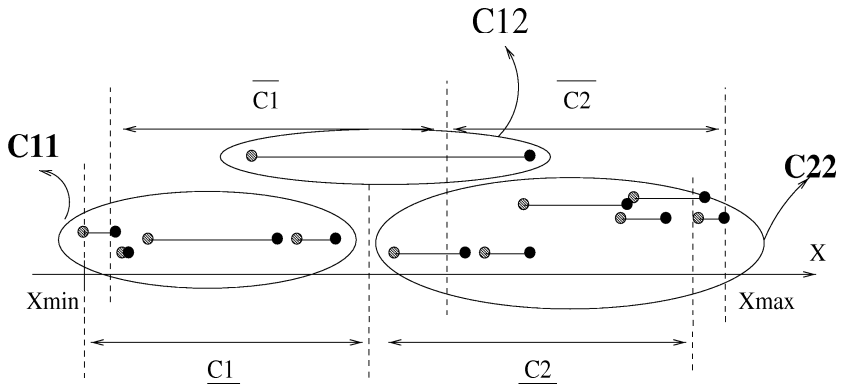


FIG. 2.2: *Les classes d'intervalles issues du découpage croisé*

2.2.3 Fonctions d'appartenance associées aux classes d'intervalles

On note $f_{\underline{C_r}}$ la fonction d'appartenance associée à la classe $\underline{C_r}$, définie dans $\Re$ à valeur dans $[0, 1]$, et qui associe à chaque valeur son degré d'appartenance à la classe $\underline{C_r}$; soit :

$$\begin{aligned} f_{\underline{C_r}} &: \Re \rightarrow [0,1] \\ x &\rightarrow f_{\underline{C_r}}(x) \end{aligned}$$

Les fonctions $f_{\underline{C_r}}$ vérifient pour toutes les bornes inférieures $\underline{x_i}$ de X_{min} la relation suivante :

$$\forall\, \underline{x_i} \in X_{min} \quad \sum_{r=1}^{k_{min}} f_{\underline{C_r}}(\underline{x_i}) \;=\; 1 \tag{2.2}$$

De manière similaire, on note $f_{\overline{C_l}}$ la fonction d'appartenance $f_{\overline{C_l}}$ associée à la classe $\overline{C_l}$, définie sur $\Re$ à valeur dans $[0,1]$, qui associe à chaque valeur son degré d'appartenance à la classe $\overline{C_l}$; soit :

$$\begin{aligned} f_{\overline{C_l}} \;: \Re \;&\rightarrow [0,1] \\ x \;&\rightarrow f_{\overline{C_l}}(x) \end{aligned}$$

Les fonctions $f_{\overline{C_l}}$ vérifient pour toutes les bornes supérieures $\overline{x_i}$ de X_{max} la relation suivante :

$$\forall\, \overline{x_i} \in X_{max} \quad \sum_{l=1}^{k_{max}} f_{\overline{C_l}}(\overline{x_i}) \;=\; 1 \tag{2.3}$$

Dans le cas des fonctions semi-linéaires floues, la représentation des fonctions d'appartenance associées aux classes $\underline{C_r}$ et $\overline{C_l}$ définies en 2.2 est donnée par la figure suivante :

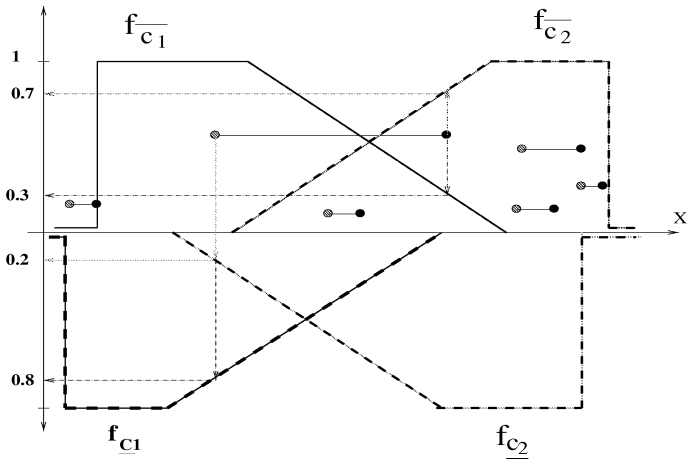


FIG. 2.3:

Connaissant les fonctions d'appartenance f_{C_r} et $f_{\overline{C_r}}$ des variables X_{min} et X_{max} il s'agit de définir les fonctions d'appartenance de la variable X produit cartésien de X_{min} et de X_{max}. Rappelant que la classe C_{rl} est définie par le croisement des classes $\underline{C_r}$ et $\overline{C_l}$, on définit $f_{C_{rl}}$ la fonction d'appartenance associée à la classe C_{rl} comme suit :

$$\begin{array}{rcl} f_{C_{rl}} : X & \rightarrow & [0,1] \\ (\underline{x},\overline{x}) & \rightarrow & f_{C_{rl}}((\underline{x},\overline{x})) = \dfrac{t\text{norme}(f_{\underline{C_r}}(\underline{x}), f_{\overline{C_l}}(\overline{x}))}{\sum_{(s,t)\in E} \ t\text{norme}(f_{\underline{C_s}}(\underline{x}), f_{\overline{C_t}}(\overline{x}))} \end{array}$$

où rappelons-le, $s \in [1..k_{min}]$, $t \in [1..k_{max}]$ désignent les indices des classes issues du découpage des variables X_{min} et X_{max} et E désigne l'ensemble des couples (s,t) d'indices définissant les croisements valides entre les classes $\underline{C_s}$ et $\overline{C_t}$.

La norme triangulaire [Schweitzer et al.61] est un opérateur d'intersection défini de $[0,1] \times [0,1]$ à valeur dans $[0,1]$ et vérifiant les propriétés suivantes :

1. T(x, y)= T(y,x) (commutativité)
2. T(x, T(y,z)) = T (T(x,y) , z) (associativité)
3. T(x,y) $\leq$ T(z,t) si $x \leq z$ et $y \leq t$ (monotonie)
4. T(x,1) = x (élément neutre 1)

La fonction d'appartenance dans le cas d'une tnorme de Zadeh

$$\begin{array}{rcl} f_{C_{rl}} : \quad X & \rightarrow & [0,1] \\ (\underline{x},\overline{x}) & \rightarrow & f_{C_{rl}}((\underline{x},\overline{x})) = \dfrac{min(f_{\underline{C_r}}(\underline{x}), f_{\overline{C_l}}(\overline{x}))}{\sum_{(s,t)\in E} \ min(f_{\underline{C_s}}(\underline{x}), f_{\overline{C_t}}(\overline{x}))} \end{array}$$

La fonction d'appartenance dans le cas d'une tnorme probabiliste

$$\begin{array}{rlll} f_{C_{rl}}: & X & \rightarrow [0,1] \\ & [\underline{x},\overline{x}] & \rightarrow f_{C_{rl}}([\underline{x},\overline{x}]) = f_{\underline{C_r}}(\underline{x}).f_{\overline{C_l}}(\overline{x}) \end{array}$$

La fonction d'appartenance dans le cas d'une tnorme Lukasiewicz

$$\begin{array}{rlll} f_{C_{rl}}: & X & \rightarrow [0,1] \\ & (\underline{x},\overline{x}) & \rightarrow f_{C_{rl}}((\underline{x},\overline{x})) = \dfrac{max(f_{\underline{C_r}}(\underline{x}) + f_{\overline{C_l}}(\overline{x}) - 1, 0)}{\sum_{(s,t)\in E} \quad max(f_{\underline{C_s}}(\underline{x}) + f_{\overline{C_t}}(\overline{x}) - 1, 0)} \end{array}$$

2.2.4 Quelques cas particuliers de codage croisé

Considérons dans ce qui suit comme opérateur d'intersection la *tnorme* probabiliste.

Le codage binaire croisé

Le codage binaire croisé est un codage croisé dont les fonctions d'appartenance associées aux classes sont des fonctions binaires. Supposant le codage binaire d'une variable X de type intervalle en quatre modalités correspondant à un découpage des classes X_{min} et X_{max} en deux classes chacune (régions de faibles et de fortes valeurs).

Les fonctions d'appartenance $f_{\underline{C_r}}, f_{\overline{C_l}}$ associées, respectivement, aux classes $\underline{C_r}, \overline{C_l}$ sont représentées dans la figure 2.4 et définies comme suit :

$$\begin{array}{rlll} f_{\underline{C_r}} & : X_{min} & \rightarrow \{0,1\} & \\ & \underline{x} & \rightarrow f_{\underline{C_r}}(\underline{x}) = 1 & si \quad \underline{x} \in \underline{C_r} \\ & & \quad\;\; = 0 & sinon \end{array}$$

$$\begin{array}{rlll} f_{\overline{C_l}} & : X_{max} & \rightarrow \{0,1\} & \\ & \overline{x} & \rightarrow f_{\overline{C_l}}(\overline{x}) = 1 & si \quad \overline{x} \in \overline{C_l} \\ & & \quad\;\; = 0 & sinon \end{array}$$

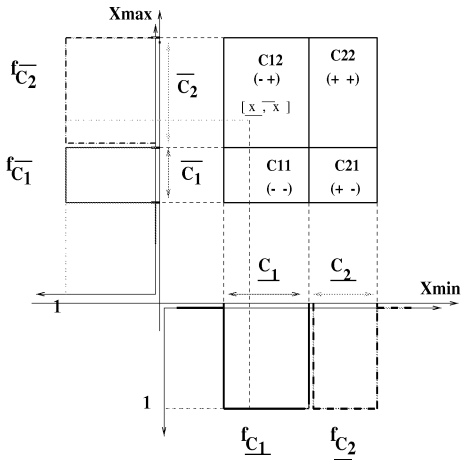


FIG. 2.4:

La définition des fonctions d'appartenance $f_{C_{rl}}$ associées aux classes est :

$$\begin{array}{lll} f_{C_{rl}} : & X & \rightarrow \{0,1\} \\ & [\underline{x},\overline{x}] & \rightarrow f_{C_{rl}}([\underline{x},\overline{x}]) = f_{\underline{C_r}}(\underline{x}).f_{\overline{C_l}}(\overline{x}) = 1 \quad si \quad \underline{x} \in \underline{C_r} \quad et \quad \overline{x} \in \overline{C_l} \\ & & \qquad\qquad\qquad\qquad\qquad\qquad = 0 \quad sinon \end{array}$$

Le codage semi-linéaire croisé

Le codage semi-linéaire croisé d'une variable de type intervalle est un codage croisé dont les fonctions d'appartenance $f_{\underline{C_r}}$ et $f_{\overline{C_r}}$ sont des fonctions semi-linéaires (figure 2.5). Considérons le découpage précédent des variables X_{min} et X_{max}. On note $\underline{c_i^c}$ le centre de la classe $\underline{C_i}$ et $\overline{c_i^c}$ le centre de la classe $\overline{C_i}$. Les fonctions semi-linéaires $f_{\underline{C_r}}$ et $f_{\overline{C_l}}$ sont représentées dans la figure 2.5 et définies comme suit :

$$\begin{array}{lll} f_{\underline{C_1}} & : X_{min} & \rightarrow [0,1] \\ & \underline{x} & \rightarrow f_{\underline{C_1}}(\underline{x}) = \dfrac{\underline{c_2^c} - \underline{x}}{\underline{c_2^c} - \underline{c_1^c}} \text{ si } \underline{x} \in [\underline{c_1^c}, \underline{c_2^c}] \\ & & \qquad\qquad = 1 \text{ si } \underline{x} \leq \underline{c_1^c} \\ & & \qquad\qquad = 0 \text{ si } \underline{x} \geq \underline{c_2^c} \end{array}$$

$$
\begin{aligned}
f_{\underline{C_2}} \; : X_{min} \rightarrow \quad & [0,1] \\
\underline{x} \rightarrow \quad & f_{\underline{C_2}}(\underline{x}) = \frac{\underline{x} - \underline{c_1^c}}{\underline{c_2^c} - \underline{c_1^c}} \text{ si } \underline{x} \in [\underline{c_1^c}, \underline{c_2^c}] \\
= \quad & 1 \text{ si } \underline{x} \geq \underline{c_2^c} \\
= \quad & 0 \text{ si } \underline{x} \leq \underline{c_1^c}
\end{aligned}
$$

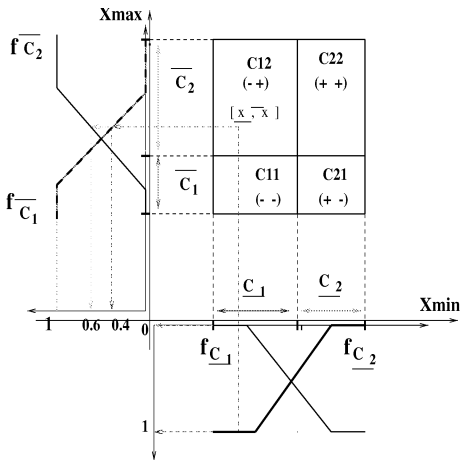


FIG. 2.5:

On définit de même les fonctions d'appartenance $f_{\overline{C_1}}$ et $f_{\overline{C_2}}$:

$$
\begin{aligned}
f_{\overline{C_1}} \; : X_{max} \quad & \rightarrow [0,1] \\
\underline{x} \quad & \rightarrow f_{\overline{C_1}}(\overline{x}) = \frac{\overline{c_2^c} - \overline{x}}{\overline{c_2^c} - \overline{c_1^c}} \text{ si } \overline{x} \in [\overline{c_1^c}, \overline{c_2^c}] \\
& = 1 \text{ si } \overline{x} \leq \overline{c_1^c} \\
& = 0 \text{ si } \overline{x} \geq \overline{c_2^c}
\end{aligned}
$$

$$
\begin{aligned}
f_{\overline{C_2}} \; : X_{max} \rightarrow \quad & [0,1] \\
\overline{x} \rightarrow \quad & f_{\overline{C_2}}(\overline{x}) = \frac{\overline{x} - \overline{c_1^c}}{\overline{c_2^c} - \overline{c_1^c}} \text{ si } \overline{x} \in [\overline{c_1^c}, \overline{c_2^c}] \\
= \quad & 1 \text{ si } \overline{x} \geq \overline{c_2^c} \\
= \quad & 0 \text{ si } \overline{x} \leq \overline{c_1^c}
\end{aligned}
$$

En considérant la *tnorme* probabiliste, la fonction semi-linéaire floue associée, par exemple, à la classe C_{11} est alors définie comme suit :

$$\begin{array}{rll} f_{C_{11}} : & X & \to [0,1] \\ & [\underline{x},\overline{x}] & \to f_{C_{11}}([\underline{x},\overline{x}]) = f_{\underline{C_1}}(\underline{x}) f_{\overline{C_1}}(\overline{x}) \end{array}$$

2.2.5 Exemple

Considérons la variable X donnant la description de la hauteur au garrot de 13 races de chiens :

Races de Chiens	Hauteur-Garrot
Caniche	[20,35]
Chihuahua	[16,20]
Pékinois	[20,25]
Basset	[26,40]
Pointer	[60,65]
Setter	[53,62]
Labrador	[54,62]
Lévrier	[55,76]
Mastiff	75
Ber-Allema	[58,65]
Dog-Allema	[76,80]
Doberman	[68,70]
Saint-bern	70

TAB. 2.2:

Représentation d'une distribution d'intervalles

Une première technique de visualisation (2.6) consiste à visualiser les intervalles observés dans un plan. La première dimension correspond à la variable X elle-même, elle permet de visualiser le positionnement des intervalles les uns par rapport aux autres. La deuxième dimension correspond à l'amplitude de variation des intervalles.

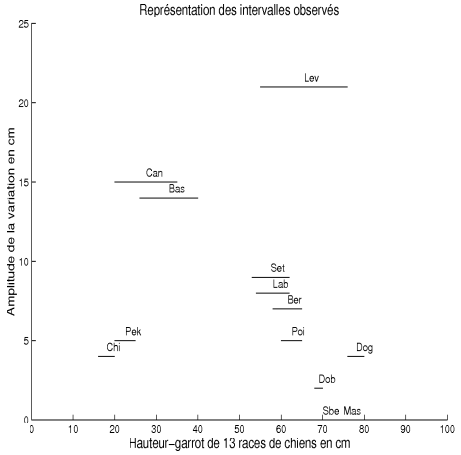


FIG. 2.6:

Dans ce type de représentation, plus un intervalle est proche du premier axe, moins son amplitude est élevée ; au contraire plus un intervalle est loin du premier axe, plus son amplitude est importante. Les intervalles triviaux ($\underline{x} = \overline{x}$) sont représentés par des points situés sur le premier axe.

Un deuxième type de visualisation de la distribution des observations de type intervalle consiste à représenter l'ensemble des intervalles dans un plan défini par les variables X_{min} et X_{max}. Chaque intervalle est alors représenté par un point situé dans le demi-plan supérieur. Ce type de représentation permet, essentiellement, de visualiser les classes d'intervalles issues du codage croisé.

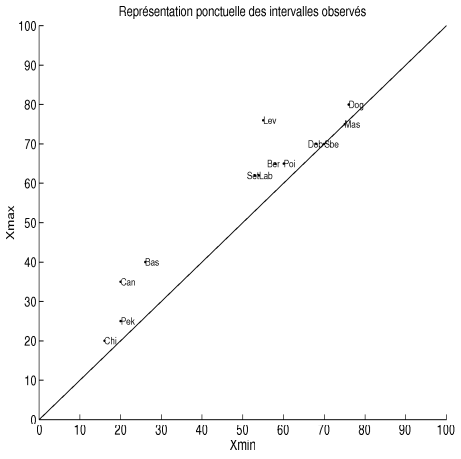


FIG. 2.7:

On distingue clairement deux principaux groupements d'intervalles.

La décomposition de la variable X

Soit X_{min} et X_{max} les deux variables issues de la décomposition horizontale de la variable *Hauteur-garrot*. Avant le découpage des variables X_{min} et X_{max}, étudions la distribution de ces variables. Pour cela on peut observer séparément les histogrammes des fréquences (voir 2.8) des variables X_{min} et X_{max}.

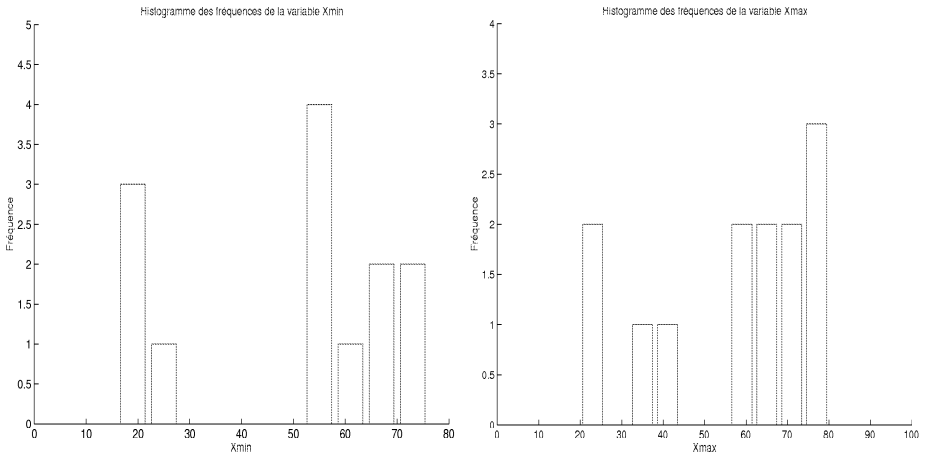


FIG. 2.8:

Les histogrammes montrent un découpage naturel de chaque variable en deux classes de valeurs (valeurs faibles, valeurs fortes).

Le codage linéaire croisé de la variable Hauteur-garrot

On propose dans cet exemple d'utiliser pour le codage des variables numériques X_{min} et X_{max} un codage linéaire en deux classes. On définit les fonctions d'appartenance $f_{\underline{C_1}}$ et $f_{\underline{C_2}}$ associées à la variable X_{min} comme suit :

$$\begin{aligned} f_{\underline{C_1}}(\underline{x}) &= \frac{76 - \underline{x}}{76 - 16} \\ f_{\underline{C_2}}(\underline{x}) &= \frac{\underline{x} - 16}{76 - 16} \end{aligned}$$

De même on définit les fonctions d'appartenance associées aux deux classes de la variable X_{max} comme suit :

$$\begin{aligned} f_{\overline{C_1}}(\overline{x}) &= \frac{80-\overline{x}}{80-20} \\ f_{\overline{C_2}}(\overline{x}) &= \frac{\overline{x}-20}{80-20} \end{aligned}$$

Le codage linéaire en deux classes (modalités) de chacune des variables numériques X_{min} et X_{max} est alors :

Races de Chiens	X_{min}		X_{max}	
	$\underline{C_1}$	$\underline{C_2}$	$\overline{C_1}$	$\overline{C_2}$
	-	+	-	+
Caniche	0.9333	0.0667	0.7500	0.2500
Chihuahua	1.0000	0	1.0000	0
Pékinois	0.9333	0.0667	0.9167	0.0833
Basset	0.8333	0.1667	0.6667	0.3333
Pointer	0.2667	0.7333	0.2500	0.7500
Setter	0.3833	0.6167	0.3000	0.7000
Labrador	0.3667	0.6333	0.3000	0.7000
Lévrier	0.3500	0.6500	0.0667	0.9333
Mastiff	0.0167	0.9833	0.0833	0.9167
Ber-allema	0.3000	0.7000	0.2500	0.7500
Dog-allema	0	1.0000	0	1.0000
Doberman	0.1333	0.8667	0.1667	0.8333
Saint-bern	0.1000	0.9000	0.1667	0.8333

TAB. 2.3: *Codage linéaire des variables X_{min} et X_{max}*

En utilisant, par exemple, la *tnorme* probabiliste, le codage linéaire croisé de la variable X en quatre classes (quatre modalités) est alors :

Races de Chiens	Hauteur-Garrot			
	C_{11}	C_{21}	C_{12}	C_{22}
	- -	+ -	- +	+ +
Caniche	0.70	0.05	0.2333	0.0167
Chihuahua	1.00	0	0	0
Pékinois	0.8556	0.0611	0.0778	0.0056
Basset	0.5556	0.1111	0.2778	0.0556
Pointer	0.0667	0.1833	0.20	0.5500
Setter	0.1150	0.1850	0.2683	0.4317
Labrador	0.11	0.19	0.2567	0.4433
Lévrier	0.0233	0.0433	0.3367	0.6067
Mastiff	0.0014	0.0819	0.0153	0.9014
Ber-allema	0.0750	0.1750	0.2250	0.5250
Dog-allema	0	0	0	1.0000
Doberman	0.0222	0.1444	0.1111	0.7222
Saint-bern	0.0167	0.15	0.0833	0.75

TAB. 2.4: *Le codage linéaire croisé de la variable X en 4 modalités*

où, par exemple, le codage de la race *caniche* pour la modalité C_{21} (+ -) est obtenu comme suit :

$$f_{C_{21}}([20,35]) \quad = \quad f_{\underline{C_2}}(20).f_{\overline{C_1}}(35) = 0.0067 * 0.75 = 0.050$$

On vérifie que la somme des codages sur chaque ligne est bien égale à 1.

2.3 Codage par sommets d'une variable de type intervalle

Dans le codage par sommets, la transformation des variables intervalles en variables numériques est fondée sur la décomposition des objets (*décomposition verticale*). Chaque ligne, donnant la description d'un objet par au moins une variable de type intervalle, est décomposée en autant de lignes que de sommets le décrivant.

Après la décomposition de l'ensemble des objets, les variables de type intervalle devenues de type numérique sont codées à l'aide des techniques classiques de codage. Le codage de chaque objet est alors défini par le codage de l'ensemble des sommets qui le constituent. Nous allons détailler chacune de ces étapes, puis illustrer le codage par sommets à l'aide d'un exemple.

2.3.1 Décomposition des objets

Soit la matrice suivante donnant la description d'un ensemble d'objets $H_1, \ldots, H_m$ par p variables de type intervalle $X_1, \ldots, X_p$ et r variables $Y_1, \ldots, Y_r$ d'autres types.

$$\begin{array}{c} \\ \begin{pmatrix} H_1 \\ \ldots \\ H_m \end{pmatrix} \end{array} = \begin{array}{c} \begin{array}{cccc|ccc} X_1 & \ldots & X_p & & Y_1 & \ldots & Y_r \end{array} \\ \begin{pmatrix} [\underline{x_{11}}, \overline{x_{11}}] & .. & [\underline{x_{1p}}, \overline{x_{1p}}] & | & y_{11} & .. & y_{1r} \\ & .. & & | & & .. & \\ [\underline{x_{m1}}, \overline{x_{m1}}] & .. & [\underline{x_{mp}}, \overline{x_{mp}}] & | & y_{m1} & .. & y_{mr} \end{pmatrix} \end{array}$$

On note q_i le nombre d'intervalles non triviaux décrivant l'objet H_i. En s'inspirant des résultats établis au chapitre précédent, on propose de décrire chaque objet H_i par l'ensemble de ses sommets.

Dans une première étape, on propose de décomposer chaque ligne donnant la description d'un objet H_i en autant de lignes (2^{q_i}) que de sommets le décrivant. On note S_j^i le $j^{ème}$ sommet de l'objet H_i. Les valeurs des variables Y_k décrivant l'objet H_i sont communes à tous ses sommets. Ainsi, la variation entre les sommets d'un même objet est uniquement introduite par les intervalles. La description de l'objet H_i à travers l'ensemble de ses sommets est donnée par la matrice suivante :

$$H_i = \begin{pmatrix} H_1^i \\ \ldots \\ S_{2^{q_i}}^i \end{pmatrix} = \begin{array}{c} \begin{array}{ccc|ccc} X_1 & \ldots & X_p & Y_1 & \ldots & Y_r \end{array} \\ \begin{pmatrix} \underline{x_{i1}} & \ldots & \underline{x_{ip}} & y_{i1} & \ldots & y_{ir} \\ \cdots & \ldots & \cdots & y_{i1} & \ldots & y_{ir} \\ \overline{x_{i1}} & \ldots & \overline{x_{ip}} & y_{i1} & \ldots & y_{ir} \end{pmatrix} \end{array}$$

La matrice obtenue suite à la décomposition de l'ensemble des objets est :

$$\begin{pmatrix} H_1 \\ \vdots \\ H_m \end{pmatrix} = \begin{pmatrix} S_1^1 \\ \dots \\ S_{2^{q_1}}^1 \\ \vdots \\ S_1^m \\ \dots \\ S_{2^{q_m}}^m \end{pmatrix} = \begin{pmatrix} \begin{matrix} X_1 & \dots & X_p & Y_1 & \dots & Y_r \end{matrix} \\ \begin{bmatrix} \underline{x_{11}} & \cdots & \underline{x_{1p}} & y_{11} & \dots & y_{1r} \\ \vdots & \ddots & \vdots & & & \\ \overline{x_{11}} & \cdots & \overline{x_{1p}} & y_{11} & \dots & y_{1r} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \underline{x_{m1}} & \cdots & \underline{x_{mp}} & y_{m1} & \dots & y_{mr} \\ \\ \vdots & \ddots & \vdots & & & \\ \overline{x_{m1}} & \cdots & \overline{x_{mp}} & y_{m1} & \dots & y_{mr} \end{bmatrix} \end{pmatrix}$$

où les variables X_j sont de type numérique.

2.3.2 Codage des sommets

Après l'étape de décomposition, on procède au codage des variables X_j (numériques) en utilisant une technique de codage appropriée. Soit $C_1^j, \dots, C_{k_j}^j$ les k_j classes (modalités) issues du découpage de la variable X_j et $\varphi_{C_1}^j, \dots, \varphi_{C_{k_j}}^j$ les fonctions d'appartenance associées aux modalités. Le codage de l'objet H_i, pour la variable X_j, est défini par le codage de l'ensemble des sommets de H_i pour cette variable. Soit :

$$H_i = \begin{pmatrix} H_1^i \\ \dots \\ S_{2^{q_i}}^i \end{pmatrix} = \begin{pmatrix} \begin{matrix} & X_j & \\ C_1^j & \dots & C_{k_j}^j \end{matrix} \\ \begin{matrix} \varphi_{C_1}^j(\underline{x_{ij}}) & \dots & \varphi_{C_{k_j}}^j(\underline{x_{ij}}) \\ \dots & \ddots & \dots \\ \varphi_{C_1}^j(\overline{x_{ij}}) & \dots & \varphi_{C_{k_j}}^1(\overline{x_{ij}}) \end{matrix} \end{pmatrix}$$

2.3.3 Exemple

Considérons l'exemple suivant donnant la description de 3 races de chiens par les variables *Hauteur-Garrot* (Hg), *Poids* (Ps) et *Fonction* (Fc) :

$$\begin{pmatrix} Caniche \\ Mastiff \\ S-Bernard \end{pmatrix} = \begin{array}{c} \begin{matrix} Hg & Ps & Fc \end{matrix} \\ \begin{pmatrix} [20,35] & [15,25] & Compagnie \\ 75 & 100 & Utile \\ 70 & [55,80] & Utile \end{pmatrix} \end{array}$$

1- Étape décomposition

La description des sommets des 3 objets après la décomposition est fournie dans le tableau suivant :

$$\begin{pmatrix} Can \\ Can \\ Can \\ Can \\ Mas \\ Sber \\ Sber \end{pmatrix} = \begin{array}{c} \begin{matrix} Hg & Ps & Fc \end{matrix} \\ \begin{pmatrix} 20 & 15 & Compagnie \\ 35 & 15 & Compagnie \\ 20 & 25 & Compagnie \\ 35 & 25 & Compagnie \\ 75 & 100 & Utile \\ 70 & 55 & Utile \\ 70 & 80 & Utile \end{pmatrix} \end{array}$$

2 Étape codage

On propose d'utiliser un codage linéaire en deux classes pour chacune des variables *Hauteur-Garrot*, *Poids*, et un codage binaire disjonctif pour la variable *Fonction* . Le codage des sommets associés aux trois races de chiens est :

Races de Chiens	*Hg*		*Ps*		*Fc*	
	Hg-	Hg+	Ps-	Ps+	Com	Uti
Can	1.0000	0	1.0000	0	1	0
Can	0.8333	0.1667	1.0000	0	1	0
Can	1.0000	0	0.8125	0.1875	1	0
Can	0.8333	0.1667	0.8125	0.1875	1	0
Mas	0	1.0000	0	1.0000	0	1
Sber	0.0833	0.9167	0.5625	0.4375	0	1
Sber	0.0833	0.9167	0.2500	0.7500	0	1

TAB. 2.5: *Codage par sommets des variables Hauteur-Garrot, Poids et Fonction*

2.4 Codage des variables intervalles sans décomposition

Contrairement aux codages croisé et par sommets qui se basent sur la transformation des variables intervalles en des variables numériques, le présent codage se base sur l'extension des outils classiques de codage à des variables intervalles.

On s'intéresse, dans une première partie, à l'extension des techniques classiques de découpage d'une variable en classes à des variables de type intervalle. Après l'introduction de nouvelles notions telles que : recouvrement entre intervalles, effectif et fréquence d'une classe d'intervalle, densité d'une classe d'intervalle, etc., on propose une généralisation des notions d'histogramme et de fonction de répartition à des distributions d'intervalles. À l'issue de ces résultats, on propose une généralisation des techniques de découpage d'une variable en classes d'effectifs égaux, ou à partir d'un histogramme, à des variables de type intervalle.

Après le découpage d'une variable intervalle en classes, on s'intéresse, dans une deuxième partie, aux fonctions d'appartenance associées aux classes qui permettent d'associer à chaque intervalle observé son degré d'appartenance à chacune des classes d'intervalles. Après l'étude des principales mesures de dissimilarité et de distance entre des intervalles, ainsi que des propriétés que doit requérir une fonction d'appartenance associée à une classe d'intervalles, nous proposons une fonction d'appartenance fondée sur une mesure de distance proposée par Moore [Moore66].

2.4.1 Recouvrement, effectif et densité d'une classe d'intervalles

Recouvrement d'un intervalle

Définition

On définit le recouvrement d'un intervalle $[\underline{x_i}, \overline{x_i}]$ par un intervalle$[\underline{x_k}, \overline{x_k}]$ noté $R_{[\underline{x_i}, \overline{x_i}]}([\underline{x_k}, \overline{x_k}])$ comme la proportion de l'intervalle $[\underline{x_i}, \overline{x_i}]$ recouverte

par l'intervalle $[\underline{x_k}, \overline{x_k}]$. On note $|\,|^{\sqcup}$ l'amplitude d'un intervalle ; il est défini comme suit :

$$\begin{aligned} |\,[\underline{x_i}, \overline{x_i}]\,|^{\square} &= \overline{x_i} - \underline{x_i} \\ |\,\phi\,|^{\sqcup} &= 0 \\ |\,[\underline{x_i}, \overline{x_i}]\,|^{\square} = |\,[\underline{x_i}, \overline{x_i}[\,|^{\square} &= |\,]\underline{x_i}, \overline{x_i}]\,|^{\square} = |\,]\underline{x_i}, \overline{x_i}[\,|^{\square} \end{aligned}$$

où ϕ désigne l'intervalle vide. On rappelle la définition de l'opérateur d'intersection $\cap$:

$$\begin{aligned} [\underline{x_i}, \overline{x_i}] \cap [\underline{x_k}, \overline{x_k}] &= [max(\underline{x_i}, \underline{x_k}), min(\overline{x_i}, \overline{x_k})] \text{ si } \quad max(\underline{x_i}, \underline{x_k}) \leq min(\overline{x_i}, \overline{x_k}) \\ &= \phi \quad \text{sinon} \end{aligned}$$

Le recouvrement $R_{[\underline{x_i}, \overline{x_i}]}([\underline{x_k}, \overline{x_k}])$ est défini comme suit :

$$R_{[\underline{x_i}, \overline{x_i}]}([\underline{x_k}, \overline{x_k}]) = \begin{cases} \frac{|[\underline{x_i}, \overline{x_i}] \cap [\underline{x_k}, \overline{x_k}]|^{\square}}{|[\underline{x_i}, \overline{x_i}]|^{\square}} & \text{si } \underline{x_i} \neq \overline{x_i} \\ 1 & \text{si } \underline{x_i} = \overline{x_i} \text{ et } \underline{x_i} \in [\underline{x_k}, \overline{x_k}] \\ 0 & \text{si } \underline{x_i} = \overline{x_i} \text{ et } \underline{x_i} \notin [\underline{x_k}, \overline{x_k}] \end{cases}$$

Propriétés

a) R est non symétrique

$$R_{[\underline{x_i}, \overline{x_i}]}([\underline{x_k}, \overline{x_k}]) = r \implies R_{[\underline{x_k}, \overline{x_k}]}([\underline{x_i}, \overline{x_i}]) = r \tag{2.4}$$

b) Soit p sous-intervalles $I_1, .., I_p$ avec $I_j = [a_{j-1}, a_j[$ discrétisant l'intervalle $[\underline{x_i}, \overline{x_i}]$ (avec $a_0 = \underline{x_i}$ et $a_p = \overline{x_i}$) ; la relation suivante est vérifiée :

$$|[\underline{x_i}, \overline{x_i}]|^{\sqcup} = \sum_{j=1}^{p} |I_j|^{\sqcup} \tag{2.5}$$

Preuve

Il est simple de constater que les termes a_i s'annulent dans l'expression $\sum_{j=1}^{p} |I_j|^{\square} = (a_1 - a_0) + \ldots + (a_p - a_{p-1}) = (a_p - a_0) = |[\underline{x_i}, \overline{x_i}]|^{\square}$.

c) $0 \leq R_{[\underline{x_i},\overline{x_i}]}([\underline{x_k}, \overline{x_k}]) \leq 1$.

d) Si $[\underline{x_i}, \overline{x_i}] \subset [\underline{x_k}, \overline{x_k}]$ alors $R_{[\underline{x_i},\overline{x_i}]}([\underline{x_k}, \overline{x_k}]) = 1$.

e) Si $[\underline{x_i}, \overline{x_i}] \cap [\underline{x_k}, \overline{x_k}] = \phi$ alors $R_{[\underline{x_i},\overline{x_i}]}([\underline{x_k}, \overline{x_k}]) = 0$.

La densité de recouvrement d'une classe d'intervalles

On note $C_1, .., C_k$, k classes issues du découpage du domaine de variation $[x_{min}, x_{max}]$ d'une distribution d'intervalles $[\underline{x_1}, \overline{x_1}], \ldots, [\underline{x_m}, \overline{x_m}]$.

Définition

On définit la densité de recouvrement $D_R(C_j)$ d'une classe C_j comme étant la somme des recouvrements des intervalles $[\underline{x_i}, \overline{x_i}]$ par la classe C_j ; soit :

$$D_R(C_j) \quad = \quad \sum_{i=1}^{m} R_{[\underline{x_i},\overline{x_i}]}(C_j)$$

Afin de dégager les propriétés des densités de recouvrement des classes, étudions les propriétés des marges de la table (2.4.1) donnant les recouvrements des intervalles $[\underline{x_i}, \overline{x_i}]$ $(i = 1..m)$ par les classes C_j $(j = 1..k)$:

$R_{[\underline{x_i},\overline{x_i}]}(C_j)$	C_1	$\ldots$	C_k	$R_i(.)$
$[\underline{x_1}, \overline{x_1}]$	$R_1(1)$	$\ldots$	$R_1(k)$	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$[\underline{x_m}, \overline{x_m}]$	$R_m(1)$	$\ldots$	$R_m(k)$	1
$R_.(j)$	$D_R(C_1)$	$\ldots$	$D_R(C_k)$	m

où

$$\begin{aligned} R_i(j) &= R_{[\underline{x_i},\overline{x_i}]}(C_j) \\ R_i(.) &= \sum_{j=1}^{k} R_i(j) \\ R_.(j) &= \sum_{i=1}^{m} R_i(j) = D_R(C_j) \end{aligned}$$

$R_i(.)$ mesure de combien une observation $[\underline{x_i}, \overline{x_i}]$ est incluse dans l'ensemble des classes et $R_.(j)$ mesure de combien la classe C_j recouvre l'ensemble des observations.

Propriétés

$$\text{a)} \qquad \forall\, i = \{1..m\} \;\; R_i(.) = 1. \tag{2.6}$$

Preuve

Sachant que $C_1, \ldots, C_k$ forment une partition sur $[x_{min}, x_{max}]$, on peut écrire pour tout intervalle $[\underline{x_i}, \overline{x_i}]$:

$$\forall\, i = \{1..m\} \quad [\underline{x_i}, \overline{x_i}] = [\underline{x_i}, \overline{x_i}] \cap [x_{min}, x_{max}] = [\underline{x_i}, \overline{x_i}] \cap (\cup_{j=1}^{k} C_j)$$

En utilisant la distributivité de $\cap$ par rapport à $\cup$:

$$[\underline{x_i}, \overline{x_i}] = \cup_{j=1}^{k} ([\underline{x_i}, \overline{x_i}] \cap C_j) \tag{2.7}$$

D'autre part, pour tout couple de classes C_j, C_l la relation suivante est vérifiée :

$$([\underline{x_i}, \overline{x_i}] \cap C_j) \cap ([\underline{x_i}, \overline{x_i}] \cap C_l) = [\underline{x_i}, \overline{x_i}] \cap C_j \cap C_l = \phi \tag{2.8}$$

Ainsi, $([\underline{x_i}, \overline{x_i}] \cap C_1), \ldots, ([\underline{x_i}, \overline{x_i}] \cap C_k)$ forment bien une partition sur $[\underline{x_i}, \overline{x_i}]$. D'après la propriété 2.5, on déduit l'égalité suivante :

$$\sum_{j=1}^{k} |[\underline{x_i}, \overline{x_i}] \cap C_j|^{\sqcup} = |[\underline{x_i}, \overline{x_i}]|^{\sqcup}$$

d'où,

$$R_i(.) \quad = \quad \frac{\sum_{j=1}^{k} |[\underline{x_i}, \overline{x_i}] \cap C_j|^{\square}}{|[\underline{x_i}, \overline{x_i}]|^{\square}} = \frac{|[\underline{x_i}, \overline{x_i}]|^{\square}}{|[\underline{x_i}, \overline{x_i}]|^{\square}} = 1$$

b) Par définition, la marge en ligne $R_.(j)$ est bien la densité de recouvrement $D_R(C_j)$ de la classe C_j ; on en déduit la propriété suivante :

$$\sum_{j=1}^{k} D_R(C_j) = \sum_{j=1}^{k} R_.(j) \quad = \quad \sum_{i=1}^{m} R_i(.) = m \tag{2.9}$$

La densité de recouvrement : une généralisation de la notion d'effectif d'une classe

Montrons que dans le cas où tous les intervalles sont triviaux (se réduisent à des points), la densité de recouvrement $D_R(C_j)$ d'une classe C_j exprime bien l'effectif d'une classe. Soit :

$$\forall\, i = \{1..m\} \quad \text{si } \overline{x_i} = \underline{x_i} = x_i \quad \Rightarrow \quad D_R(C_j) = n_j$$

où n_j est l'effectif de la classe C_j (i.e. le nombre de points tombant dans C_j).

Preuve

Dans le cas où tous les intervalles $[\underline{x_i}, \overline{x_i}]$ sont des points, l'expression de la densité de recouvrement devient :

$$D_R(C_j) \quad = \quad \sum_{i=1}^{m} R_{[\underline{x_i}, \overline{x_i}]}(C_j)$$

Or, par définition $R_{[\underline{x_i}, \overline{x_i}]}(C_j)$ est égal à 1 si le point $x_i \in C_j$ et à 0 sinon ; soit,

$$D_R(C_j) \quad = \quad \sum_{i=1}^{n_j} 1 = n_j$$

où n_j est le nombre d'intervalles (points) $[x_i, x_i]$ inclus dans la classe C_j; ce qui correspond bien à l'effectif de la classe Cj. La densité de recouvrement d'une classe est bien une généralisation de la notion d'effectif d'une classe à des données intervalles.

La fréquence de recouvrement d'une classe

On définit la fréquence de recouvrement d'une classe C_j notée $F_R(C_j)$ comme suit :

$$F_R(C_j) = \frac{D_R(C_j)}{\sum_{l=1}^{k} D_R(C_l)} = \frac{D_R(C_j)}{m}$$

2.4.2 Histogramme d'une distribution d'intervalles

Discrétisation du domaine d'une distribution d'intervalles

Soit $[\underline{x_1}, \overline{x_1}], \ldots, [\underline{x_m}, \overline{x_m}]$ une distribution de m intervalles. On note x_{min}, x_{max} et a_{min}, respectivement, la plus petite valeur, la plus grande valeur et la plus petite amplitude non nulle observée ; soit :

$$\begin{aligned} x_{min} &= \min_{i=1..m} (\underline{x_i}) \\ x_{max} &= \max_{i=1..m} (\overline{x_i}) \\ a_{min} &= \min_{i=1..m} \{(\overline{x_i} - \underline{x_i}) \ / \ \text{avec} \ \ \overline{x_i} \neq \underline{x_i}\} \end{aligned}$$

On ne dispose pas actuellement de règles absolues permettant de déterminer le nombre de classes d'un histogramme. Un nombre de classes trop faible risque de faire perdre l'information de distribution ; à l'opposé, un nombre de classes trop élevé risque de faire apparaître des classes vides ou d'effectifs très faibles. Dans le cas numérique continu, la discrétisation se fait en général en considérant des classes d'égales amplitudes.

De manière similaire, on peut choisir de discrétiser le domaine $[x_{min}, x_{max}]$ de variation des intervalles en classes d'égales amplitudes. On suggère quelques recommandations pour le choix du nombre de classes, dans le cas d'une distribution d'intervalle. Sachant que chaque intervalle est comptabilisé comme une

unité dans le calcul des fréquences, il est inutile de choisir un pas de discrétisation inférieur à la plus petite amplitude a_{min} non nulle. Il est préférable de choisir un nombre de classe inférieur à $\frac{x_{max} - x_{min}}{a_{min}}$.

Histogramme d'une distribution d'intervalles

Le découpage le plus élémentaire consiste à découper le domaine $[x_{min}, x_{max}]$ en k classes d'égales amplitudes. Si la distribution des intervalles n'est pas équirépartie, en général on utilise l'histogramme des fréquences pour définir les limites des classes. On évite ainsi de construire des classes vides ou de faibles densités. Pour la construction de l'histogramme, on procède comme suit :

On discrétise d'abord le domaine de variation de la distribution d'intervalles en p intervalles $I_1, \ldots, I_p$. Pour chaque sous-intervalle $I_j = [a_{j-1}, a_j]$, on porte en ordonnée la fréquence de recouvrement $\frac{F_R(I_j)}{|I_j|^{\square}}$ divisée par l'amplitude de ce sous-intervalle. Si les intervalles Ij sont de même amplitude, on porte en ordonnée directement les fréquences de recouvrement $F_R(I_j)$.

Notons dans ce contexte, le travail proposé par Decarvalho [Carvalho92], [Carvalho95] étendant la notion d'histogramme à des objets symboliques et fondée sur la notion de potentiel de description.

Remarque Une autre approche de discrétisation consiste à visualiser tous les intervalles observés sur un axe, puis en partant de la plus petite valeur x_{min} jusqu'à la plus grande x_{max}, à marquer la limite d'une nouvelle classe chaque fois que l'on rencontre une borne inférieure ou une borne supérieure d'un intervalle. L'effectif d'une classe ainsi obtenue est le nombre d'intervalles la recouvrant. Le risque d'une telle technique est d'avoir un nombre élevé de classes de très faibles densités.

2.4.3 Fonction de répartition d'une distribution d'intervalles

Avant de définir la fonction de répartition d'une distribution d'intervalles, démontrons la propriété d'additivité des densités de recouvrement suivante :

Additivité des densités de recouvrement

Soit un intervalle quelconque $[\underline{x}, \overline{x}] \subset [x_{min}, x_{max}]$, et $I_1, \ldots, I_r$ avec $(I_j = [a_{j-1}, a_j])$ r sous-intervalles discrétisant l'intervalle $[\underline{x}, \overline{x}]$. On montre que la densité de recouvrement de $[\underline{x}, \overline{x}]$ est égale à la somme des densités de recouvrement des r sous-intervalles discrétisant $[\underline{x}, \overline{x}]$, soit :

$$D_R([\underline{x}, \overline{x}]) = \sum_{j=1}^{r} D_R(I_j)$$

Preuve

En substituant $[\underline{x}, \overline{x}]$ par son expression $I_1 \cup \ldots \cup I_r$, la densité de recouvrement s'exprime alors :

$$D_R([\underline{x}, \overline{x}]) = \sum_{i=1}^{m} \frac{|(\cup_{j=1}^{r} I_j) \cap [\underline{x_i}, \overline{x_i}]|^{\square}}{|[\underline{x_i}, \overline{x_i}]|^{\square}} \tag{2.10}$$

$$\tag{2.11}$$

En utilisant la distributivité de l'intersection par rapport à l'union, on obtient :

$$D_R([\underline{x}, \overline{x}]) = \sum_{i=1}^{m} \frac{|\cup_{j=1}^{r} (I_j \cap [\underline{x_i}, \overline{x_i}])|^{\square}}{|[\underline{x_i}, \overline{x_i}]|^{\square}}$$

D'après la propriété énoncée en 2.5, l'expression devient :

$$\begin{aligned} D_R([\underline{x}, \overline{x}]) &= \sum_{i=1}^{m} \frac{\sum_{j=1}^{r} |I_j \cap [\underline{x_i}, \overline{x_i}]|^{\square}}{|[\underline{x_i}, \overline{x_i}]|^{\square}} \\ &= \sum_{j=1}^{r} (\sum_{i=1}^{m} \frac{|(I_j \cap [\underline{x_i}, \overline{x_i}])|^{\square}}{|[\underline{x_i}, \overline{x_i}]|^{\square}}) \end{aligned}$$

d'où

$$D_R([\underline{x}, \overline{x}]) = \sum_{j=1}^{r} D_R(I_j) \tag{2.12}$$

La fonction de répartition d'une distribution d'intervalles

Définition

La définition de la fonction de répartition notée F^I d'une distribution d'intervalles se fonde sur la propriété d'additivité des densités de recouvrement définie en 2.10. Soit F^I la fonction de répartition définie sur $[x_{min}, x_{max}]$:

$$F^I(x) = \begin{cases} 0 & \text{Si } x \leq x_{min} \\ F_R([x_{min}, x]) & \text{Si } x \in [x_{min}, x_{max}] \\ 1 & \text{Si } x \geq x_{max} \end{cases} \tag{2.13}$$

Propriétés

La fonction F^I vérifie les propriétés suivantes :

a) $0 \leq F^I(x) \leq 1$

b) $\lim_{x \to x_{min}} F^I(x) \to 0$

c) $\lim_{x \to xmax} F^I(x) \to 1$

Preuve

$$\begin{aligned} \lim_{x \to x_{min}} F^I(x) &= \lim_{x \to x_{min}} F_R([x_{min}, x]) = \lim_{x \to x_{min}} \frac{D_R([x_{min}, x])}{m} \\ &= \lim_{x \to x_{min}} \frac{1}{m} \sum_{i=1}^{m} \frac{|[\underline{x_i}, \overline{x_i}] \cap [x_{min}, x]|^{\square}}{|[\underline{x_i}, \overline{x_i}]|^{\square}} \end{aligned}$$

or

$$\lim_{x \to x_{min}} [\underline{x_i}, \overline{x_i}] \cap [x_{min}, x] \to \begin{cases} [x_{min}, x_{min}] & \text{si } x_{min} \in [\underline{x_i}, \overline{x_i}] \\ \phi & \text{sinon} \end{cases}$$

d'où

$$\lim_{x \to x_{min}} F^I(x) \to \frac{1}{m} \sum_{i=1}^{m} 0 \to 0;$$

De manière similaire,

$$\begin{aligned}\lim_{x\to x_{max}} F^I(x) &= \lim_{x\to x_{max}} F_R([x_{min},x]) = \lim_{x\to x_{max}} \frac{D_R([x_{min},x])}{m} \\ &= \lim_{x\to x_{max}} \frac{1}{m}\sum_{i=1}^{m} \frac{|[\underline{x_i},\overline{x_i}]\cap[x_{min},x]|^{\sqcup}}{|[\underline{x_i},\overline{x_i}]|^{\square}}\end{aligned}$$

or

$$\forall i\in\{1..m\} \quad \lim_{x\to x_{max}} [\underline{x_i},\overline{x_i}]\cap[x_{min},x] \quad\to\quad [\underline{x_i},\overline{x_i}]$$

d'où

$$\lim_{x\to x_{max}} F^I(x_{max}) \quad\to\quad \frac{1}{m}\sum_{i=1}^{m}\frac{|[\underline{x_i},\overline{x_i}]|^{\square}}{|[\underline{x_i},\overline{x_i}]|^{\square}} \to \frac{1}{m}\sum_{i=1}^{m} 1 \to 1$$

ainsi,

$$\lim_{x\to x_{max}} F^I(x) \quad\to\quad 1$$

c) $F^I(x)$ est une fonction non décroissante.

Preuve

Soit a, b $\in [x_{min}, x_{max}]$ tels que $a \le b$

$$\begin{aligned}F^I(a) = F_R([x_{min},a]) &= \frac{D_R([x_{min},a])}{m} \\ F^I(b) = F_R([x_{min},b]) &= \frac{D_R([x_{min},b])}{m}\end{aligned}$$

d'après la propriété d'additivité définie en 2.10, on peut écrire :

$$D_R([x_{min},b]) \quad = \quad D_R([x_{min}a]) + D_R([a,b])$$

comme $D_R([a,b]) \ge 0$, alors

$$\frac{D_R([x_{min},a])}{m} \le \frac{D_R([x_{min},b])}{m} \quad\Rightarrow\quad F^I(a)\le F^I(b)$$

2.4.4 Généralisation des techniques de découpage aux variables intervalles

Soit X une variable de type intervalle observée pour m objets $H_1, \ldots, H_m$. On note $[\underline{x_i}, \overline{x_i}]$ l'observation de la variable X pour l'objet H_i.

Découpage d'une variable intervalle par histogramme

Le découpage de la variable intervalle X à l'aide de l'histogramme des fréquences consiste simplement à construire l'histogramme des fréquences de la distribution d'observations $[\underline{x_1}, \overline{x_1}], \ldots, [\underline{x_m}, \overline{x_m}]$ comme cela est indiqué en 2.4.2 puis, comme dans le cas des variables numériques, à choisir les coupures définissant les classes d'intervalles.

Découpage d'une variable intervalle en k classes d'effectifs égaux

Le problème du découpage la variable intervalle X en k classes de mêmes effectifs revient à rechercher k classes $C_1, \ldots, C_k$ (avec $C_j = [\underline{c_j}, \overline{c_j}]$), tels que :

$$\forall j = 1..k \qquad D_R(Cj) = \frac{m}{k}$$

En d'autres termes, il s'agit de déterminer les bornes supérieures des classes C_j (j=1..k-1) en connaissant la fréquence cumulée $F^I(\overline{c_j}) = \frac{j}{k}$ en ces bornes.

Pour cela, on procède par interpolation, comme dans le cas numérique. On note $I_1, .., I_p$ ($I_j = [a_{j-1}, a_j]$) les p sous-intervalles discrétisant le domaine $[x_{min}, x_{max}]$ et $F^I(x)$ la fréquence cumulée au point x. Les bornes supérieures des classes sont définies comme suit :

$$\overline{c_j} = \frac{(\frac{j}{k} - F^I(a_{r-1}))(a_r - a_{r-1})}{F^I(a_r) - F^I(a_{r-1})} + a_{r-1}$$

où l'intervalle $I_r = [a_{r-1}, a_r[$ contenant la valeur $\overline{c_j}$ est déterminé comme suit :

$$F^I(a_{r-1}) < \frac{j}{k} \leq F^I(a_r)$$

2.4.5 Mesures de dissimilarité ou de distance entre intervalles

L'objectif de cette section est de définir les fonctions d'appartenance associées aux classes d'intervalles (issues du découpage d'une variable intervalle), et qui permettent de mesurer le degré d'appartenance d'un intervalle observé à une classe d'intervalles.

Partant de l'idée qu'une classe est elle-même un intervalle, et que le degré d'appartenance d'un intervalle à la classe peut être interprété comme une mesure de proximité entre l'intervalle observé et l'intervalle définissant la classe, nous nous sommes alors intéressés aux travaux portant sur les mesures de dissimilarité et de distance entre les intervalles. La mesure de proximité entre un intervalle et une classe (déduite d'une mesure de dissimilarité ou de distance entre intervalles) définit alors le degré d'appartenance de l'intervalle à la classe en question.

Nous présentons d'abord quelques mesures usuelles de dissimilarité ou de distance définies sur les intervalles. Les caractéristiques de chaque mesure sont illustrées à l'aide d'une représentation et d'une interprétation géométrique. Nous définissons ensuite les propriétés que doit vérifier une fonction de codage d'une variable intervalle en vue d'une ACM. Ces mêmes propriétés sont étudiées pour chaque mesure de distance ou de dissimilarité présentée. Finalement, on propose une fonction d'appartenance tenant compte des propriétés du codage préalablement fixé.

Interprétations géométriques reliées aux intervalles

On présente, dans un premier temps, la représentation géométrique de quelques notions de base reliées aux intervalles : amplitude d'un intervalle, union et intersection de deux intervalles, positions particulières de deux intervalles, etc. Ces représentations géométriques nous permettent par la suite d'illustrer clairement certaines propriétés ou concepts relatifs aux mesures de distance et de dissimilarité.

1) Amplitude d'un intervalle

L'amplitude $(\overline{a} - \underline{a})$ d'un intervalle a correspond à la longueur séparant le point a de sa projection verticale sur la première bissectrice, soit :

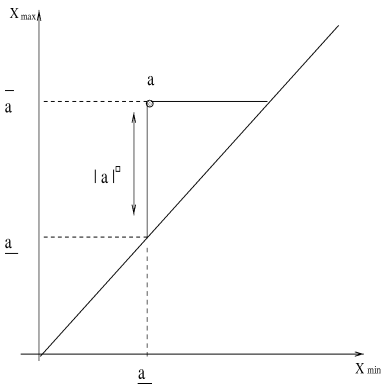
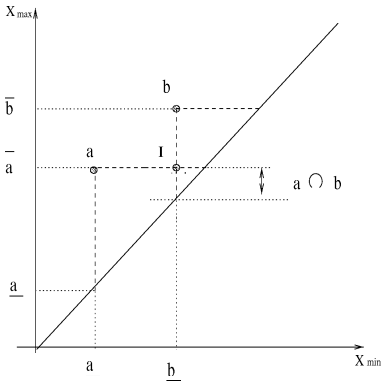


FIG. 2.9:

2) Intersection de deux intervalles

On associe à tout intervalle a un triangle isocèle défini par projection verticale et horizontale avec la première bissectrice. L'intersection $a \cap b$ de deux intervalles a et b est l'intervalle I défini par le point d'intersection entre les deux triangles associés aux intervalles a et b.



Si les deux triangles sont disjoints pas alors, les intervalles a et b ne se recouvrent pas et leur intersection est nulle.

2) Union de deux intervalles

On définit l'union de deux intervalles a et b notée $a \cup^I b$ comme suit :

$$a \cup^I b \quad = \quad [min(\underline{a}, \underline{b}), max(\overline{a}, \overline{b})]$$

elle est représentée comme suit :

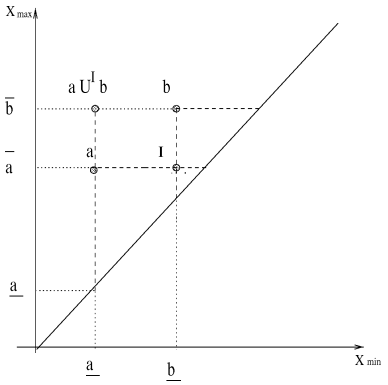
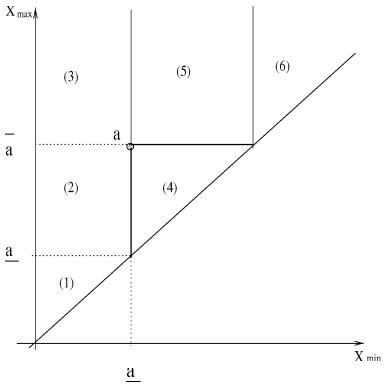


FIG. 2.10:

3) Les positions particulières de deux intervalles

On découpe le demi-plan supérieur en régions correspondant à des positions particulières des deux intervalles.



Tout intervalle $b = [\underline{b}, \overline{b}]$ appartenant à la région (i) admet une position particulière par rapport à l'intervalle a comme indiqué ci-dessous :

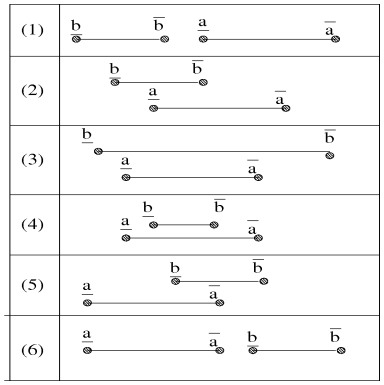


FIG. 2.11:

Les mesures de distance ou de dissimilarité entre les intervalles

Plusieurs travaux se sont intéressés aux mesures de proximité entre les intervalles. MOORE [Moore66] propose une mesure de distance euclidienne définie sur les intervalles et fondée sur l'écart entre les bornes inférieure et supérieure. GOWDA, DIDAY [Gowda et al.92], [Gowda et al.91a] proposent des mesures de similarités ou de dissimilarité entre les intervalles se fondant sur des mesures d'amplitudes, de recouvrements et de positions. Ichino [Ichino et al.94] propose une mesure de distance entre les intervalles fondée sur des opérateurs de jointure et d'union définis sur des intervalles. La distance de HAUSDORFF peut également être appliquée à des données intervalles. On présente dans ce qui suit les mesures de dissimilarité et de distance énoncées et on étudie leurs caractéristiques.

Distance euclidienne de MOORE

Définition

Soit $a = [\underline{a}, \overline{a}]$, $b = [\underline{b}, \overline{b}]$ deux intervalles, R. MOORE [Moore66] propose une mesure de distance euclidienne notée d_M et définie comme suit :

$$d_M(a,b) \quad = \quad \sqrt{(\underline{a} - \underline{b})^2 + (\overline{a} - \overline{b})^2}$$

Interprétation géométrique

La distance d_M définie précédemment peut être représentée comme suit :

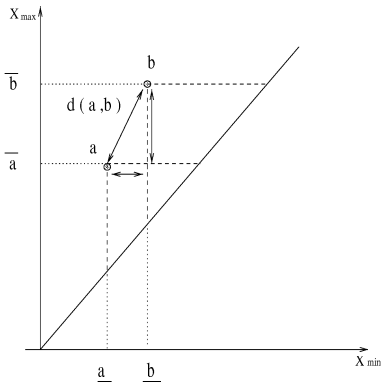


FIG. 2.12:

Remarquons que l'ensemble des intervalles (l'ensemble des couples ordonnés) constitue un sous-ensemble de $\Re^2$ (demi-plan supérieur), et que la distance d_M n'est autre que la distance euclidienne définie sur $\Re^2$. D'autre part, les intervalles $b = [\underline{b}, \overline{b}]$ dont la distance $d_M(a, b)$ de a est égale à r, sont situés sur le cercle de centre l'intervalle a et de rayon r.

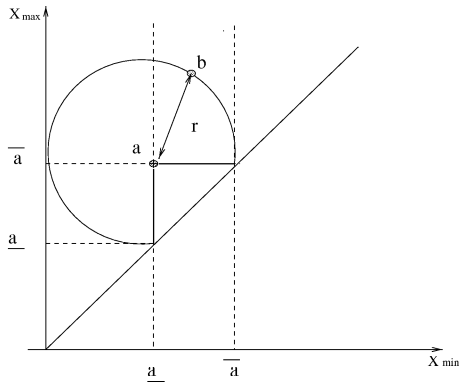


FIG. 2.13:

Mesures de dissimilarité de Gowda et Diday

Définition

La mesure de dissimilarité proposée par Gowda [Gowda et al.91a] se fonde sur trois indices. Le premier indice noté D_p permet de mesurer la position relative de deux intervalles a et b en se fondant sur l'écart entre leurs bornes inférieures $|\underline{b} - \underline{a}|$ rapporté à l'amplitude du domaine $[x_{min}, x_{max}]$. Le domaine $[x_{min}, x_{max}]$ inclut tout intervalle observé.

$$D_p(a,b) \quad = \quad \frac{|\underline{a} - \underline{b}|}{|[x_{min}, x_{max}]|^{\square}}$$

Le second indice D_s permet de mesurer la différence d'amplitude des deux intervalles rapportée à l'union des deux intervalles.

$$D_s(a,b) \quad = \quad \frac{||a|^{\sqcup} - |b|^{\sqcup}|}{|a \cup^I b|^{\square}}$$

Le troisième indice D_c mesure l'amplitude des régions de a et de b non communes rapportée à l'union des intervalles.

$$D_c(a,b) \quad = \quad \frac{|a|^{\square} + |b|^{\square} - 2|a \cap b|^{\square}}{|a \cup^I b|^{\square}}$$

La mesure de dissimilarité proposée par Gowda notée d_G est définie comme la somme des trois indices D_p, D_s et D_c ; soit :

$$d_G(a,b) = D_p(a,b) + D_s(a,b) + D_c(a,b)$$

Interprétation géométrique

Les numérateurs des trois indices sont représentés comme suit :

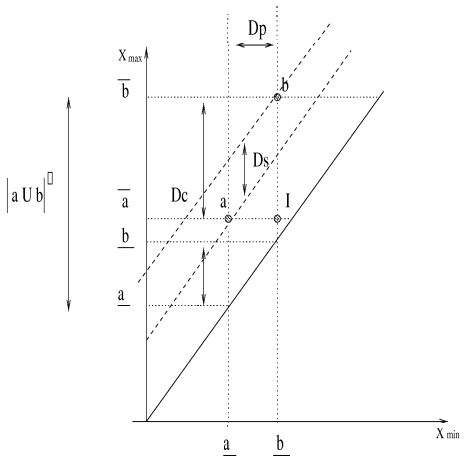


FIG. 2.14:

En étudiant séparément les cas ou les intervalles a et b se recouvrent ou pas, la dissimilarité d_G proposée par GOWDA peut s'exprimer comme suit :

	$d_G(a,b)$
$a \cap^I b = \phi$	$\frac{\lvert\underline{b}-\underline{a}\rvert}{\lvert x_{min}, x_{max}\rvert^{\square}} + \frac{2\ max(\lvert a\rvert^{\square}, \lvert b\rvert^{\square})}{\lvert a \cup^I b\rvert^{\square}}$
$a \cap^I b \neq \phi$	$\frac{\lvert\underline{b}-\underline{a}\rvert}{\lvert x_{min}, x_{max}\rvert^{\square}} - \frac{2\ min(\lvert a\rvert^{\square}, \lvert b\rvert^{\square})}{\lvert a \cup^I b\rvert^{\square}} + 2$

Distance d'ICHINO

Définition

Soit a et b deux intervalles. La distance proposée par ICHINO [Ichino et al.94], notée d_I^γ, est fondée sur les deux opérateurs d'union $a \cup^I b$ et d'intersection $a \cap^I b$ définis sur les intervalles. La distance d_I^γ est définie comme suit :

$$d_I^\gamma(a,b) = \left|a \cup^I b\right|^\square - \left|a \cap^I b\right|^\square + \gamma(2\left|a \cap^I b\right|^\square - |a|^\square - |b|^\square) \quad (2.14)$$

où $0 \leq \gamma \leq 0.5$.

En étudiant séparément les cas où les intervalles sont disjoints, se recouvrent avec ou sans inclusion, la distance d_I^γ proposée par ICHINO s'exprime comme suit :

$d_I^\gamma(a,b)$	$\gamma = 0$	$0 < \gamma < 0.5$	$\gamma = 0.5$
$a \cap b = \phi$	$\lvert a \cup^I b\rvert^\square$	$\geq\ d_I^\gamma(a,b) \geq$	$\lvert b_c - a_c\rvert$
$a \cap b \neq \phi$ et $a \not\subset b$	$2\lvert b_c - a_c\rvert$	$\geq\ d_I^\gamma(a,b)\ \geq$	$\lvert b_c - a_c\rvert$
$a \subset b$	$2((b_c - a_c) + (\underline{b} - \underline{a}))$	$\geq\ d_I^\gamma(a,b) \geq$	$((b_c - a_c) + (\underline{b} - \underline{a}))$

où a_c et b_c sont les centres des intervalles a et b.

Preuve

Étudions les cas où les intervalles sont disjoints, se recouvrent avec ou sans inclusion. Considérons, ce qui ne restreint pas la généralité, que $\overline{a} \leq \overline{b}$.

1) $a \cap b = \phi$

La distance d_I^γ se réduit à :

$$d_I^\gamma \quad = \quad |a \cup^I b|^\square - \gamma(|a|^\square + |b|^\square)$$

si $0 \leq \gamma \leq 0.5$, alors :

$$\left|a \cup^I b\right|^\square - \frac{|a|^\square + |b|^\square}{2} \quad \leq \quad d_I^\gamma(a,b) \leq \left|a \cup^I b\right|^\square$$

on vérifie aisément que $\left|a \cup^I b\right|^\square - \frac{|a|^\square + |b|^\square}{2} = b_c - a_c$; ainsi, la distance d_I varie comme suit :

$$d_I^{0.5}(a,b) = |b_c - a_c| \quad \leq \quad d_I^\gamma \leq |a \cup^I b|^\square = d_I^0(a,b)$$

2) $a \cap b \neq \phi$ **et** $a \not\subset b$

Pour $0 \leq \gamma \leq 0.5$, d_I^γ est bornée par :

$$\begin{aligned} d_I^\gamma(a,b) \quad &\leq \quad \left|a \cup^I b\right|^\square - |a \cap b|^\square = d_I^0(a,b) \\ &\geq \quad \left|a \cup^I b\right|^\square - |a \cap b|^\square + (|a \cap b|^\square - \frac{|a|^\square}{2} - \frac{|b|^\square}{2}) = d_I^{0.5}(a,b) \end{aligned}$$

or

$\left|a \cup^I b\right|^\square - |a \cap b|^\square = 2(b_c - a_c)$ et

$|a \cup^I b|^\square - |a \cap b|^\square + |a \cap b|^\square - \frac{|a|^\sqcup}{2} - \frac{|b|^\sqcup}{2} = (b_c - a_c)$, ainsi :

$$d_I^{0.5}(a,b) = |b_c - a_c| \quad \leq \quad d_I^\gamma(a,b) \leq 2|b_c - a_c| = d_I^0(a,b)$$

$$d_I^{0.5}(a,b) = |b_c - a_c| \quad \leq \quad d_I^\gamma \leq |a \cup^I b|^\square = d_I^0(a,b)$$

3) $a \subset b$

La distance d_I^γ est réduite à :

$$d_I^\gamma \quad = \quad (1-\gamma)(|b|^\square - |a|^\square)$$

pour $0 \leq \gamma \leq 0.5$, la distance varie dans :

$$d_I^{0.5}(a,b) = \frac{|b|^\square - |a|^\square}{2} \quad \leq \quad d_I^\gamma \leq |b|^\square - |a|^\square = d_I^0(a,b)$$

sachant que $|b|^\square - |a|^\square = 2\left((b_c - a_c) - (\underline{b} - \underline{a})\right)$, alors

$$d_I^{0.5}(a,b) = \left((b_c - a_c) - (\underline{b} - \underline{a})\right) \quad \leq \quad d_I^\gamma \leq 2\left((b_c - a_c) - (\underline{b} - \underline{a})\right) = d_I^0(a,b)$$

Distance de Hausdorff appliquée aux intervalles

Définition

Étant donné deux ensembles de points A et B, la distance de Hausdorff $d_H(A,B)$ entre A et B est égale à r si et seulement si chaque point de A est à une distance de moins de r d'au moins un point de B, et si, réciproquement, chaque point de B est à une distance de moins de r d'au moins un point de A. Soit :

$$d_H(A,B) \quad = \quad max\left(\max_{\beta \in B}(\min_{\alpha \in A} d(\alpha,\beta)) \ , \ \max_{\alpha \in A}(\min_{\beta \in B} d(\alpha,\beta))\right)$$

Sachant qu'un intervalle $a = [\underline{a}, \overline{a}]$ est un ensemble de points, la distance de Hausdorff entre des intervalles est définie comme suit :

$$d_H(a,b) \quad = \quad max(|\underline{b} - \underline{a}|, |\overline{b} - \overline{a}|)$$

2.4.6 Propriétés des fonctions d'appartenance associées aux classes d'intervalles

Étant donné des classes d'intervalles associées à une variable de type intervalle, notre objectif est de définir une fonction d'appartenance, fondée sur une mesure de dissimilarité ou de distance, qui associe à chaque observation son degré d'appartenance à chacune des classes d'intervalles. Pour cela, nous allons d'abord définir les propriétés que doit vérifier la fonction d'appartenance associée à une classe d'intervalles. Ces propriétés sont ensuite étudiées pour chacune des mesures de dissimilarité ou de distance présentées.

Degré d'appartenance d'un intervalle à une classe d'intervalles

Le degré d'appartenance d'un intervalle a à une classe C mesure de combien la région de valeurs représentée par l'intervalle a est caractéristique de celle représentée par la classe C_j. On note φ_C la fonction d'appartenance associée à la classe C, elle définit pour chaque intervalle $a = [\underline{a}, \overline{a}]$ son degré d'appartenance à la classe C.

Propriétés requises d'une fonction d'appartenance associée à une classe d'intervalles

Cas 1 : Considérons le cas de deux intervalles a et b **non inclus** dans la classe C et dont l'intersection avec la classe C est **non vide**.

1. *Si les deux intervalles* **a** *et* **b** *sont d'égale amplitude et ont des intersections avec la classe* **C** *de même amplitude, alors ils admettent un même degré d'appartenance à la classe* **C**.

$$\forall\, a\,, b \text{ tels que } |a|^{\sqcup} = |b|^{\sqcup} \text{ et } |a \cap C|^{\sqcup} = |b \cap C|^{\sqcup} \;\Rightarrow\; \varphi_C(a) = \varphi_C(b)$$

2. *Si les deux intervalles sont d'amplitude différente et ont des intersections avec la classe* **C** *d'égale amplitude, alors le degré d'appartenance le plus élevé est celui de l'intervalle de plus faible amplitude.*

$$\forall\, a\,, b \text{ tels que } |a \cap C|^{\square} = |b \cap C|^{\square} \text{ et } |a|^{\square} > |b|^{\square} \;\Rightarrow\; \varphi_C(a) < \varphi_C(b)$$

3. *Si les deux intervalles sont de même amplitude et ont des intersections avec la classe* **C** *de différente amplitude, alors le degré d'appartenance le plus élevé est celui de l'intervalle dont l'amplitude de l'intersection avec la classe est la plus élevée.*

$$\forall\, a\,, b \text{ tels que } |a|^{\square} = |b|^{\square} \text{ et } |a \cap C|^{\square} > |b \cap C|^{\square} \;\Rightarrow\; \varphi_C(a) > \varphi_C(b)$$

4. *Soit deux classes C_i et C_j de différente amplitude. Si l'intersection de l'intervalle* **a** *avec la classe C_i est d'égale amplitude avec l'intersection de* **a** *avec la classe C_j, alors le degré d'appartenance de l'intervalle* **a** *le plus élevé est celui associé à la classe de plus faible amplitude.*

$$\forall\, a\,, b \text{ tels que } |a \cap C_i|^{\sqcup} = |a \cap C_j|^{\sqcup} \text{ et } |C_i|^{\sqcup} < |C_j|^{\sqcup} \;\Rightarrow\; \varphi_{C_i}(a) > \varphi_{C_j}(a)$$

Cas 2 : Considérons le cas de deux intervalles a et b **inclus** dans une classe C.

1. *Si les intervalles* **a** *et* **b** *sont de même amplitude et sont situés symétriquement par rapport au centre de la classe, alors ils ont le même degré d'appartenance à la classe. Soit :*

$$\forall\, a \subset C\,, b \subset C \text{ tels que } |a|^{\sqcup} = |b|^{\sqcup} \text{ et } |a_c - C_c| = |b_c - C_c| \;\Rightarrow\; \varphi_C(a) = \varphi_C(b)$$

2. *Si les deux intervalles* **a** *et* **b** *sont de même amplitude, alors le degré d'appartenance le plus élevé est celui de l'intervalle dont le centre est le plus proche du centre de la classe. Soit :*

$$\forall\, a \subset C\,, b \subset C \text{ tels que } |a|^{\square} = |b|^{\square} \text{ et } \; |a_c - C_c| \leq |b_c - C_c| \;\; \Rightarrow \;\; \varphi_C(a) \geq \varphi_C(b)$$

3. *Soit deux intervalles* **a** *et* **b** *tels que* **a** *est inclus dans* **b**, *alors le degré d'appartenance de l'intervalle* **b** *à la classe* **C** *est supérieur à celui de l'intervalle* **a**.

$$\forall\, a \subset b \subset C \;\; \Rightarrow \;\; \varphi_C(a) \leq \varphi_C(b)$$

Étude des propriétés d'une fonction d'appartenance fondée sur la distance de Moore

On assimile le degré d'appartenance d'un intervalle **a** à une classe d'intervalles **C** comme une mesure de proximité entre l'intervalle **a** et l'intervalle représentant la classe **C**. On propose d'étudier la validité des propriétés définies au début de ce paragraphe (2.4.6) dans le cas d'une fonction d'appartenance ou d'une mesure de similarité fondée sur la distance d_M. Vérifions la validité des propriétés énoncées précédemment.

Propriété 1 (cas 1)

Définissons, tout d'abord, l'ordre total suivant entre les intervalles $a = [\underline{a}, \overline{a}]$ et $b = [\underline{b}, \overline{b}]$:

$$a \leq b \;\; \Leftrightarrow \;\; \begin{cases} \overline{a} < \overline{b} \\ \text{ou} \\ \overline{a} = \overline{b} \text{ et } \underline{a} \leq \underline{b} \end{cases}$$

Supposons, ce qui ne restreint pas la généralité, que $a \leq C \leq b$. On peut alors établir les égalités suivantes :

$$\underline{c} - \underline{a} = |a|^{\sqcup} - |a \cap C|^{\sqcup} \quad (2.15)$$
$$\overline{c} - \overline{a} = |C|^{\sqcap} - |a \cap C|^{\sqcap} \quad (2.16)$$

Or comme $(|a|^{\sqcup} = |b|^{\sqcup})$ et $(|a \cap C|^{\sqcup} = |b \cap C|^{\sqcup})$, alors les équations 2.15 et 2.16 s'expriment comme suit :

$$\begin{aligned} \underline{c} - \underline{a} &= |b|^{\sqcap} - |b \cap C|^{\sqcap} = \overline{b} - \overline{c} \\ \overline{c} - \overline{a} &= |C|^{\sqcup} - |b \cap C|^{\sqcup} = \underline{c} - \underline{b} \end{aligned}$$

d'où

$$d_M(a, C) = \sqrt{(\underline{a} - \underline{c})^2 + (\overline{a} - \overline{c})^2} = \sqrt{(\underline{b} - \underline{c})^2 + (\overline{b} - \overline{c})^2} = d_M(b, C)$$

Ainsi, les intervalles a et b sont bien situés à une même distance d_M par rapport à la classe C, ils ont donc un même degré de similarité (d'appartenance) avec la classe C.

Propriété 2 (cas 1)

Pour cela, supposons $a \leq C \leq b$, les équations 2.15 et 2.16 s'expriment comme suit :

$$\begin{aligned} \underline{c} - \underline{a} &= |a|^{\sqcap} - |a \cap C|^{\sqcap} > |b|^{\sqcap} - |b \cap C|^{\sqcap} = \overline{b} - \overline{c} \\ \overline{c} - \overline{a} &= |C|^{\sqcap} - |a \cap C|^{\sqcap} = |C|^{\sqcap} - |b \cap C|^{\sqcap} = \underline{c} - \underline{b} \end{aligned}$$

d'où

$$d_M(a, C) = \sqrt{(\underline{a} - \underline{c})^2 + (\overline{a} - \overline{c})^2} > \sqrt{(\underline{b} - \underline{c})^2 + (\overline{b} - \overline{c})^2} = d_M(b, C)$$

ainsi, l'intervalle a est plus léloigné (moins similaire), selon d_M de C que l'est l'intervalle b. La propriété 2 est donc vérifiée.

Propriété 3 (cas 1)

En supposant $a \leq C \leq b$, les équations 2.15 et 2.16 s'expriment comme suit :

$$\begin{aligned} \underline{c} - \underline{a} &= |a|^{\sqcap} - |a \cap C|^{\sqcap} < |b|^{\sqcap} - |b \cap C|^{\sqsubset} \\ \overline{c} - \overline{a} &= |C|^{\sqcap} - |a \cap C|^{\sqcap} < |C|^{\sqcap} - |b \cap C|^{\sqcap} \end{aligned}$$

d'où

$$d_M(a, C) = \sqrt{(\underline{a} - \underline{c})^2 + (\overline{a} - \overline{c})^2} \quad < \sqrt{(\underline{b} - \underline{c})^2 + (\overline{b} - \overline{c})^2} < d_M(b, C)$$

ainsi, l'intervalle a est plus proche (plus similaire), selon d_M de C que l'est l'intervalle b. La propriété 3 est donc vérifiée.

Propriété 4 (cas 1)

Supposons l'ordre suivant $C_i \leq a \leq C_j$.

$$\begin{aligned} \underline{a} - \underline{c_i} = |C_i|^{\sqcap} - |a \cap C_i|^{\sqcap} &< |C_j|^{\sqcap} - |a \cap C_i|^{\sqcap} = \overline{c_j} - \overline{a} \\ \overline{a} - \overline{c_i} = |a|^{\sqcup} - |a \cap C_i|^{\sqcup} &= |a|^{\sqcup} - |a \cap C_j|^{\sqcup} = \underline{c_j} - \underline{a} \end{aligned}$$

d'où

$$d_M(a, C_i) = \sqrt{(\underline{a} - \underline{c_i})^2 + (\overline{a} - \overline{c_i})^2} \quad < \sqrt{(\underline{a} - \underline{c_j})^2 + (\overline{a} - \overline{c_j})^2} = d_M(b, C_j)$$

La propriété 4 est bien vérifiée.

Étudions maintenant les propriétés dans le cas où les intervalles a et b sont inclus dans la classe C.

Propriété 1 (cas 2)

Supposons l'ordre suivant $a \leq b \leq C$. D'après les caractéristiques des intervalles a et b on peut écrire les expressions suivantes :

$$[\underline{a}, \overline{a}] = [c_c - (\Delta + |a|^{\square}), c_c - \Delta]$$

$$[\underline{b}, \overline{b}] = [c_c + \Delta, c_c + (\Delta + |b|^{\square})]$$

ainsi,

$$\begin{aligned} \underline{a} - \underline{c} &= c_c - (\Delta + |a|^{\sqcup}) - \underline{c} \\ \overline{c} - \overline{a} &= \overline{c} - c_c + \Delta \end{aligned}$$

Sachant que $c_c - \underline{c} = \overline{c} - c_c$, alors on peut réécrire les expressions précédentes comme suit, puisque $\underline{b} = \underline{c_c} + \Delta$, $\overline{b} = \underline{c} + \Delta + |b|^{\square}$

$$\begin{aligned} \underline{a} - \underline{c} &= \overline{c} - c_c - (\Delta + |b|^{\square}) = \overline{c} - \overline{b} \\ \overline{c} - \overline{a} &= c_c - \underline{c} + \Delta = \underline{b} - \underline{c} \end{aligned}$$

d'où

$$d_M(a, C) = d_M(b, C)$$

Propriété 2 (cas 2)

Considérons un intervalle a centré et inclus dans la classe C. On pose :

$$\overline{c} - \overline{a} = \underline{a} - \underline{c} = l$$

ainsi,

$$d_M(a, C) = \sqrt{2\, l^2}$$

soit $b = [\underline{a} + \Delta, \overline{a} + \Delta]$ avec $(-l \leq \Delta \leq l)$ un intervalle non centré dans C et de même amplitude que a. La distance $d_M(b, C)$ est alors :

$$\begin{aligned} d_M(b,C) &= \sqrt{(l+\Delta)^2+(l-\Delta)^2} \\ &= \sqrt{2\,l^2+2\Delta^2} \;>\; d_M(a,C) \end{aligned}$$

Ainsi, plus un intervalle inclus dans la classe est centré, plus il est similaire à la classe. La propriété 2 (cas 2) est bien vérifiée.

De même, on vérifie aisément la propriété 3 (cas 2).

Étude des propriétés d'une fonction d'appartenance fondée sur la mesure de dissimilarité de GOWDA et DIDAY

Étudions à travers quelques configurations la non validité de certaines propriétés énoncées précédemment.

Configuration 1

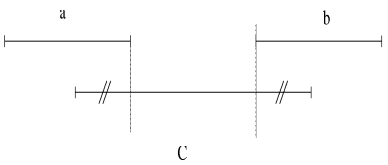


FIG. 2.15:

où les intervalles a et b de même amplitude recouvrent de la même amplitude la classe C. Calculons les dissimilarités d_G entre les intervalles a et b et la classe C.

$$\begin{aligned} d_G(a,C) &= \frac{|\underline{c}-\underline{a}|}{|[\underline{x_{min}},\overline{x_{max}}]|^{\square}} - \frac{2Min(|a|^{\square},|C|^{\square})}{|a\cup^I C|^{\square}} + 2 \\ d_G(b,C) &= \frac{|\underline{c}-\underline{b}|}{|[\underline{x_{min}},\overline{x_{max}}]|^{\square}} - \frac{2\;Min(|b|^{\square},|C|^{\square})}{|b\cup^I C|^{\square}} + 2 \end{aligned}$$

Comme $\dfrac{|b|^{\square}}{|b\cup^I C|^{\square}} = \dfrac{|a|^{\square}}{|a\cup^I C|^{\square}}$ et $|\underline{c}-\underline{b}| > |\underline{c}-\underline{a}|$, alors $d_G(a,C) < d_G(b,C)$

Ainsi, la propriété 1 (cas 1) n'est pas vérifiée.

Configuration 2

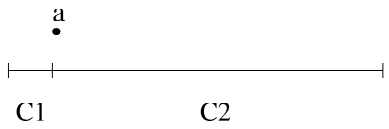


FIG. 2.16:

où $|C_1|^{\square} < |C_2|^{\square}$.

Calculons les dissimilarités de d_G entre l'intervalle a et chacune des classes C_1 et C_2.

$$\begin{aligned} d_G(a, C_1) &= \frac{|C_1|^{\square}}{|[\underline{x_{min}}, \overline{x_{max}}]|^{\square}} + 2 \\ d_G(a, C_2) &= 2 \end{aligned}$$

où U est l'amplitude du domaine $[x_{min}, x_{max}]$. On voit que quelle que soit l'amplitude (non nulle) de la classe C_1, la dissimilarité $d_G(a, C_1)$ est toujours supérieure à $d_G(a, C_2)$. Ainsi, la propriété 4 (cas 1) n'est pas vérifiée.

Configuration 3

Considérons deux intervalles a et b de même amplitude inclus dans la classe C et symétriques par rapport au centre de la classe comme suit :

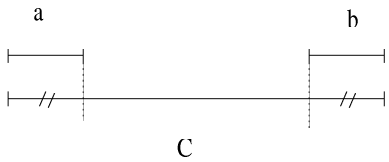


FIG. 2.17:

Les dissimilarités des intervalles a et b avec la classe C sont alors :

$$d_G(a, C) = 2 - \frac{2|a|^{\square}}{|C|^{\square}}$$

$$d_G(b,C) = \frac{\underline{b}-\underline{c}}{|[\underline{x_{min}},\overline{x_{max}}]|^{\square}} + 2 - \frac{|b|^{\sqcup}}{|C|^{\square}}$$

Comme, d'une part, $|a|^{\sqcup} = |b|^{\sqcup}$ et $(\underline{b}-\underline{c}) > 0$, alors la dissimilarité $d_G(a,C)$ est inférieure à $d_G(b,C)$. La propriété 1 (cas 2) n'est donc pas vérifiée.

Étude des propriétés d'une fonction d'appartenance fondée sur la distance d'ICHINO

Reconsidérons la configuration 1 précédente et montrons, par exemple, pour $\gamma = 0.5$, la non validité de la propriété 1 (cas 1).

$$\begin{aligned} d_I^{0.5}(a,C) &= |C_c - a_c| \\ d_I^{0.5}(b,C) &= |C_c - b_c| \end{aligned}$$

Comme les intervalles a et b sont situés symétriquement par rapport au centre de la classe C, alors $|C_c - a_c| = |C_c - b_c|$, donc $d_I^{0.5}(a,C) = d_I^{0.5}(b,C)$. La propriété 1 (cas 1) est vérifiée.

Considérons la configuration 3 et calculons la distance $d_I^{0.5}$ entre l'intervalle a, b et la classe :

$$\begin{aligned} d_I^{0.5}(a,C) &= (|C|^{\square} - |a|^{\square})/2 \\ d_I^{0.5}(b,C) &= (|C|^{\square} - |b|^{\square})/2 \end{aligned}$$

Comme a et b sont symétriquement situés par rapport au centre de la classe C, alors $d_I^{0.5}(a,C) = d_I^{0.5}(b,C)$. La propriété 1 (cas 2) est vérifiée.

Étude des propriétés d'une fonction d'appartenance fondée sur la distance de Hausdorff

Étudions le cas de figure suivant :

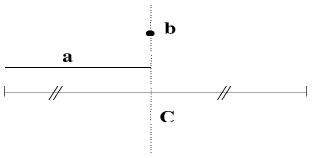


FIG. 2.18:

On utilise la distance de Hausdorff pour calculer la distance des intervalles a et b à la classe C :

$$\begin{aligned} d_H(a,C) &= |\overline{c}-\overline{a}| = \overline{c}-C_c \\ d_H(b,C) &= |\overline{c}-\overline{b}| = \overline{c}-C_c \end{aligned}$$

Ainsi, $d_H(a,C) = d_H(b,C)$; l'intervalle b est considéré aussi proche de la classe que l'intervalle a. La propriété 3 (cas 2) n'est donc pas vérifiée.

Configuration 2

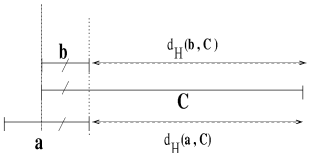


FIG. 2.19:

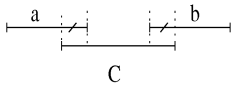
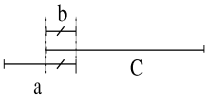
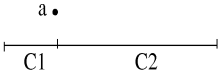
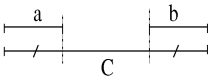
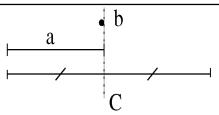
Les intervalles a et b ont des intersections de la même amplitude avec la classe C ; de plus $|a|^{\square} > |b|^{\square}$. Calculons la distance d_H entre les intervalles a, b et la classe C.

$$d_H(a,C) = |\overline{c}-\overline{a}|$$

$$d_H(b, C) \;=\; |\overline{c} - \overline{b}|$$

Comme $|\overline{c} - \overline{a}| = |\overline{c} - \overline{b}|$, alors $d_H(a, C) = d_H(b, C)$. Or l'amplitude de l'intervalle a est supérieure à celle de l'intervalle b. La propriété 2 (cas 1) n'est pas vérifiée.

Un jeu de configurations est utilisé pour illustrer les ressemblances et les différences entre les mesures de distance et de dissimilarité abordées précédement. Les deux symboles + et - marquent la cohérence, respectivement, la non cohérence du résultat avec les propriétés (définies en 2.4.6) requises pour une fonction d'appartenance. Les résultats sont récapitulés dans le tableau ci-dessous.

d_M	d_G	d_I	d_H	Configurations
$d_M(a,C) = d_M(b,C)$ +	$d_G(a,C) < d_G(b,C)$ -	$d_I(a,C) = d_I(b,C)$ +	$d_H(a,C) = d_H(b,C)$ +	
$d_M(a,C) > d_M(b,C)$ +	$d_G(a,C) > d_G(b,C)$ +	$d_I(a,C) > d_I(b,C)$ +	$d_H(a,C) = d_H(b,C)$ -	
$d_M(a,C1) < d_M(a\,C2)$ +	$d_G(a,C1) > d_G(a\,C2)$ -	$d_I(a,C1) < d_I(a\,C2)$ +	$d_H(a,C1) < d_H(a\,C2)$ +	
$d_M(a,C) = d_M(b,C)$ +	$d_G(a,C) < d_G(b,C)$ -	$d_I(a,C) = d_I(b,C)$ +	$d_H(a,C) = d_H(b,C)$ +	
$d_M(a,C) < d_M(b,C)$ +	$d_G(a,C) < d_G(b,C)$ +	$d_I(a,C) < d_I(b,C)$ +	$d_H(a,C) = d_H(b,C)$ -	

2.4.7 Fonction d'appartenance fondée sur la distance de Moore

D'après les résultats précédemment établits, nous proposons de définir un codage flou fondée sur la distance proposée par Moore et qui vérifie les propriétés préalablement fixées. On définit la fonction d'appartenance φ_j comme suit :

$$\begin{array}{rrl} \varphi_j : & X \rightarrow & [0,1] \\ & [\underline{x_i}, \overline{x_i}] \rightarrow & \varphi_j([\underline{x_i}, \overline{x_i}]) = 1 - \dfrac{d_M([\underline{x_i}, \overline{x_i}], C_j)}{\sum_{j=1}^{k} d_M([\underline{x_i}, \overline{x_i}], C_j)} \end{array}$$

2.4.8 Comparaison des trois techniques de codage d'une variable de type intervalle

Si l'objectif de l'Acm consiste à visualiser la variation intrinsèque à chaque objet, le codage par sommets est approprié. Dans un codage par sommets toutes les techniques de découpages et les fonctions d'appartenance classiques restent applicables. Chaque variable intervalle est décomposée, comme dans le cas quantitatif, en modalités chacune représentant une région de valeurs (faibles, moyennes, fortes, etc.). L'information de variation intrinsèque à chaque objet est déterminée à partir de ses sommets. Les classes, dans le codage par sommets, sont définies à partir des bornes des intervalles en tant que des observations indépendantes, ceci risque d'engendrer la construction de classes creuses, au sens des observations de type intervalle. Par ailleurs, si le nombre de variables de type intervalle est important le codage par sommets devient coûteux, il est alors préférable d'adopter un codage croisé ou sans décomposition.

Les techniques de découpages et les fonctions d'appartenance classiques restent également utilisables dans le codage croisé. Une modalité issue du codage croisé caractérise non seulement une région de valeurs mais également un niveau de variation. Par exemple, une modalité représente l'ensemble des intervalles situés dans la régions $[0, 15]$ du domaine de la variable et dont les

amplitudes varient entre $[1, 6]$. Le codage croisé ne permet pas la visualisation de la variation inhérente à chaque objet, cette information peut être retrouvée par l'interprétation des proximités des objets aux différentes modalités de la variable.

Finalement, si l'objectif du codage consiste, d'une part, à construire des classes d'effectifs égaux (au sens des observations de type intervalle), à l'aide d'un histogramme, ou d'une fonction de répartition, d'autre part, à utiliser une fonction d'appartenance vérifiant des propriétés particulières rattachées aux intervalles, alors le codage sans décomposition est le codage approprié. Les modalités obtenues, comme dans le cas quantitatif, décrivent des régions de valeurs formant une partition sur le domaine de la variable codée. L'interprétation des positions des objets par rapport aux modalités dépendra de la fonction d'appartenance utilisée.

Chapitre 3

Applications

3.1 Application en reconnaissances de visages

La reconnaissance automatique de visages fait l'objet d'un intérêt croissant ces dernières années, en particulier dans le cadre d'application de restriction et de sécurisation d'accès (bâtiments, réseaux informatiques) et de la surveillance de la vigilance (conducteurs). Le processus de reconnaissance de visages comporte trois phases successives : la description, la classification et l'identification.

La phase de description des visages consiste à extraire les caractéristiques intrinsèques du visage. On distingue deux techniques classiques de description. La technique *géométrique* [Kanade73] consiste à décrire chaque visage par un ensemble de paramètres mesurant la position relative et la taille des principaux éléments composant le visage (yeux, nez, bouche etc.). Dans la deuxième technique dite *globale* [Turk et al.91], chaque visage représenté par une image à 256×256 pixels est décrit par un vecteur à 256^2 composantes ; chaque composante mesure l'intensité ou le niveau de gris d'un pixel particulier de l'image.

La phase de classification revient en général à l'étude de la typologie des visages précédement décrits. Pour cela, on utilise l'analyse en composantes principales qu'on retrouve souvent sous le terme de *méthode de Karhunen-Loève* [Kirby et al.90]. Cette phase permet, d'une part, de déterminer les principaux groupes de visages ainsi que les descripteurs caractérisant chaque groupe ; d'autre part, elle assure une fonction de compression d'images, puisque les images de visages sont décrites en sortie dans un espace de dimension réduite.

La phase d'identification revient à comparer le jeu de paramètres d'un nouveau visage avec ceux contenus dans la base des visages, décrits et analysés dans les phases précédentes. Une mesure de distance est effectuée entre le visage a identifié et les visages de la base. L'image de la base donnant la meilleure correspondance permet d'identifier la personne, si toutefois la distance n'est pas trop élevée.

3.1.1 Présentation de l'application

Cette application concerne un système de reconnaissance de visages fondé sur l'utilisation de l'approche géométrique. Ce système s'inscrit dans le cadre du projet AMIBE [Leroy et al.96], qui se propose d'expérimenter une interface multimodale homme-machine intégrant le son et l'image pour un nombre limité d'utilisateurs. L'utilisateur d'une telle interface est d'abord identifié par une carte magnétique personnelle puis, tout au long de la transaction, l'utilisateur est filmé de face et son identité est vérifiée en utilisant les modes paroles et images.

3.1.2 Description des données

Dans cette application, la base de données des visages est constituée d'acquisitions temporelles des visages de neuf personnes de sexe masculin. Ces acquisitions sont faites en studio ; les personnes sont de face sur fond uniforme avec des conditions d'éclairage identiques. Nous disposons pour chacune des neuf personnes de données de trois séquences d'images. Chaque visage est identifié par des distances (en nombre de pixels) calculées entre des points caractéristiques du visage (figure 3.1).

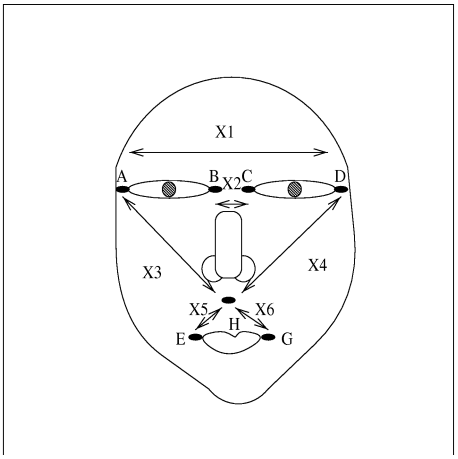


FIG. 3.1: *Les descripteurs d'un visage*

Du fait de la mobilité de la tête, les positions des points ainsi que les distances caractérisant le visage varient lors des différentes prises de vues. Afin de prendre en compte cette variation, chaque séquence d'images est décrite par des données de type intervalle. Les données intervalles sont obtenues en prenant la plus petite et la plus grande valeur observée au sein d'une même séquence d'images.

Ainsi, la base de données des visages obtenue (tableau 3.1) donne la description de 27 séquences d'images de visages à raison de 3 séquences par personne. Les séquences sont décrites par 6 variables de type intervalle prenant en compte la variation des images d'une personne lors des différentes prises de vues.

	AD	BC	AH	DH	EH	GH
FRA1	[155.00,157.00]	[58.00,61.01]	[100.45,103.28]	[105.00,107.30]	[61.40,65.73]	[64.20,67.80]
FRA2	[154.00,160.01]	[57.00,64.00]	[101.98,105.55]	[104.35,107.30]	[60.88,63.03]	[62.94,66.47]
FRA3	[154.01,161.00]	[57.00,63.00]	[99.36,105.65]	[101.04,109.04]	[60.96,65.60]	[60.42,66.40]
HUS1	[168.86,172.84]	[58.55,63.39]	[102.83,106.53]	[122.38,124.52]	[56.73,61.07]	[60.44,64.54]
HUS2	[169.85,175.03]	[60.21,64.38]	[102.94,108.71]	[120.24,124.52]	[56.73,62.37]	[60.44,66.84]
HUS3	[168.76,175.15]	[61.40,63.51]	[104.35,107.45]	[120.93,125.18]	[57.20,61.72]	[58.14,67.08]
INC1	[155.26,160.45]	[53.15,60.21]	[95.88,98.49]	[91.68,94.37]	[62.48,66.22]	[58.90,63.13]
INC2	[156.26,161.31]	[51.09,60.07]	[95.77,99.36]	[91.21,96.83]	[54.92,64.20]	[54.41,61.55]
INC3	[154.47,160.31]	[55.08,59.03]	[93.54,98.98]	[90.43,96.43]	[59.03,65.86]	[55.97,65.80]
ISA1	[164.00,168.00]	[55.01,60.03]	[120.28,123.04]	[117.52,121.02]	[54.38,57.45]	[50.80,53.25]
ISA2	[163.00,170.00]	[54.04,59.00]	[118.80,123.04]	[116.67,120.24]	[55.47,58.67]	[52.43,55.23]
ISA3	[164.01,169.01]	[55.00,59.01]	[117.38,123.11]	[116.67,122.43]	[52.80,58.31]	[52.20,55.47]
JPL1	[167.11,171.19]	[61.03,65.01]	[118.23,121.82]	[108.30,111.20]	[63.89,67.88]	[57.28,60.83]
JPL2	[169.14,173.18]	[60.07,65.07]	[118.85,120.88]	[108.98,113.17]	[62.63,69.07]	[57.38,61.62]
JPL3	[169.03,170.11]	[59.01,65.01]	[115.88,121.38]	[110.34,112.49]	[61.72,68.25]	[59.46,62.94]
KHA1	[149.34,155.54]	[54.15,59.14]	[111.95,115.75]	[105.36,111.07]	[54.20,58.14]	[48.27,50.61]
KHA2	[149.34,155.32]	[52.04,58.22]	[111.20,113.22]	[105.36,111.07]	[53.71,58.14]	[49.41,52.80]
KHA3	[150.33,157.26]	[52.09,60.21]	[109.04,112.70]	[104.74,111.07]	[55.47,60.03]	[49.20,53.41]
LOT1	[152.64,157.62]	[51.35,56.22]	[116.73,119.67]	[114.62,117.41]	[55.44,59.55]	[53.01,56.60]
LOT2	[154.64,157.62]	[52.24,56.32]	[117.52,119.67]	[114.28,117.41]	[57.63,60.61]	[54.41,57.98]
LOT3	[154.83,157.81]	[50.36,55.23]	[117.59,119.75]	[114.04,116.83]	[56.64,61.07]	[55.23,57.80]
PHI1	[163.08,167.07]	[66.03,68.07]	[115.26,119.60]	[116.10,121.02]	[60.96,65.30]	[57.01,59.82]
PHI2	[164.00,168.03]	[65.03,68.12]	[114.55,119.60]	[115.26,120.97]	[60.96,67.27]	[55.32,61.52]
PHI3	[161.01,167.00]	[64.07,69.01]	[116.67,118.79]	[114.59,118.83]	[61.52,68.68]	[56.57,60.11]
ROM1	[167.15,171.24]	[64.07,68.07]	[123.75,126.59]	[122.92,126.37]	[51.22,54.64]	[49.65,53.71]
ROM2	[168.15,172.14]	[63.13,68.07]	[122.33,127.29]	[124.08,127.14]	[50.22,57.14]	[49.93,56.94]
ROM3	[167.11,171.19]	[63.13,68.03]	[121.62,126.57]	[122.58,127.78]	[49.41,57.28]	[50.99,60.46]

TAB. 3.1: *Les données visages*

3.1.3 Résultats dans le cadre de la méthode des sommets

On applique la *méthode des sommets* au tableau de données décrit en 3.1. Les valeurs propres ainsi que les pourcentages d'inerties figurent dans le tableau 3.2. Nous constatons que les 3 premières composantes principales restituent 83.34% de l'inertie totale.

	Méthode des sommets		
Numéro	Valeurs propres	% d'inertie	Cumul
1	2.5602	42.67	42.67
2	1.7983	29.97	72.64
3	0.6422	10.70	83.34
4	0.4758	7.93	91.27
5	0.3351	5.58	96.85
6	0.1883	3.14	99.99

TAB. 3.2: *Valeurs propres et pourcentage d'inertie*

Description et représentation factorielle des visages

La description factorielle des séquences de visages, à un niveau $\alpha = 0$, dans l'espace défini par les 4 premières composantes principales, est donnée par le tableau 3.3.

	Méthode des sommets			
	CP1	CP2	CP3	CP4
FRA1	[-2.66, -1.61]	[0.27, 1.57]	[-0.29, 1.00]	[-0.52, 0.64]
FRA2	[-2.49, -1.03]	[-0.11, 1.61]	[-0.25, 1.01]	[-0.60, 1.21]
FRA3	[-2.99, -0.81]	[-0.40, 1.88]	[-0.88, 1.20]	[-0.87, 1.35]
HUS1	[-0.24, 1.10]	[0.39, 2.05]	[0.64, 2.13]	[-1.13, 0.43]
HUS2	[-0.40, 1.41]	[0.56, 2.65]	[0.29, 2.32]	[-1.22, 0.61]
HUS3	[-0.24, 1.42]	[0.43, 2.52]	[0.27, 2.17]	[-0.99, 0.53]
INC1	[-3.77, -2.29]	[-0.67, 1.23]	[-0.80, 0.69]	[-0.68, 1.22]
INC2	[-3.66, -1.35]	[-2.05, 0.92]	[-0.88, 1.83]	[-0.91, 1.92]
INC3	[-4.02, -1.86]	[-1.20, 1.41]	[-1.01, 1.50]	[-0.55, 1.56]
ISA1	[0.80, 2.01]	[-1.83, -0.46]	[-0.58, 0.58]	[-1.10, 0.34]
ISA2	[0.37, 1.86]	[-1.71, -0.08]	[-0.64, 0.73]	[-1.48, 0.13]
ISA3	[0.41, 2.11]	[-1.84, -0.12]	[-0.58, 1.20]	[-1.35, 0.33]
JPL1	[-0.36, 0.92]	[0.54, 2.03]	[-1.81, -0.43]	[-1.05, 0.36]
JPL2	[-0.34, 1.17]	[0.48, 2.37]	[-1.85, -0.07]	[-1.42, 0.34]
JPL3	[-0.52, 0.93]	[0.50, 2.28]	[-1.56, 0.25]	[-1.47, 0.34]
KHA1	[-1.18, 0.39]	[-3.08, -1.46]	[-1.19, 0.26]	[-0.19, 1.47]
KHA2	[-1.46, 0.15]	[-3.17, -1.32]	[-0.93, 0.61]	[-0.54, 1.33]
KHA3	[-1.71, 0.25]	[-2.95, -0.72]	[-1.25, 0.57]	[-0.75, 1.57]
LOT1	[-0.74, 0.61]	[-2.51, -0.87]	[-0.81, 0.61]	[-1.39, 0.17]
LOT2	[-0.69, 0.40]	[-1.94, -0.62]	[-0.80, 0.33]	[-1.39, -0.12]
LOT3	[-0.82, 0.34]	[-2.12, -0.70]	[-0.77, 0.52]	[-1.70, -0.26]
PHI1	[0.22, 1.51]	[0.56, 1.84]	[-1.40, -0.08]	[-0.12, 1.05]
PHI2	[-0.09, 1.66]	[0.33, 2.29]	[-1.81, 0.22]	[-0.53, 1.15]
PHI3	[-0.25, 1.38]	[0.25, 2.25]	[-2.01, -0.12]	[-0.58, 1.25]
ROM1	[2.19, 3.45]	[-1.20, 0.29]	[-0.51, 0.81]	[0.09, 1.48]
ROM2	[1.85, 3.63]	[-1.30, 0.97]	[-0.83, 1.36]	[-0.46, 1.50]
ROM3	[1.48, 3.57]	[-1.33, 1.31]	[-0.79, 1.79]	[-0.60, 1.60]

TAB. 3.3: *Les quatre premières composantes principales de type intervalle*

Rappelons qu'une description factorielle de niveau α signifie que les coordonnées de type intervalle ont été obtenues en prenant en compte les coordonnées factorielles des sommets dont les contributions relatives sont supérieures ou égales au seuil α. Les figures 3.2, 3.3 fournissent la représentation factorielle des 27 séquences de visages dans les plans factoriels (PC_1, PC_2) et (PC_1, PC_3) de niveau $\alpha = 0$.

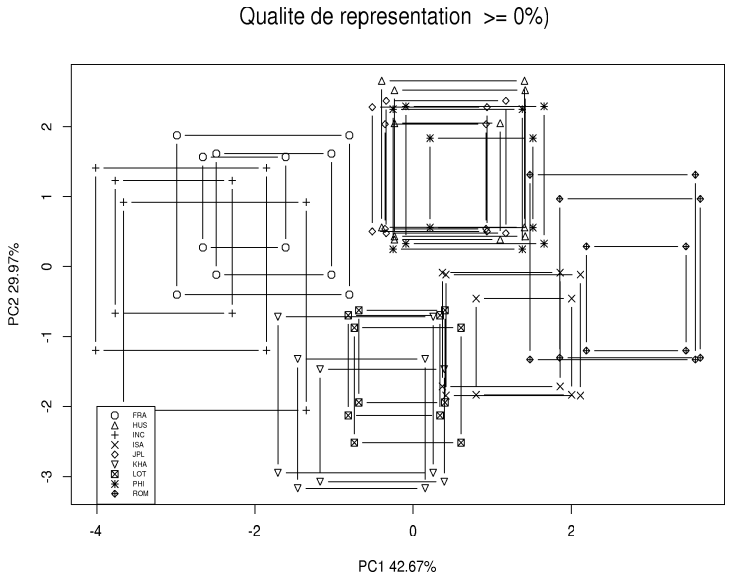


FIG. 3.2: *Premier plan factoriel de niveau 0*

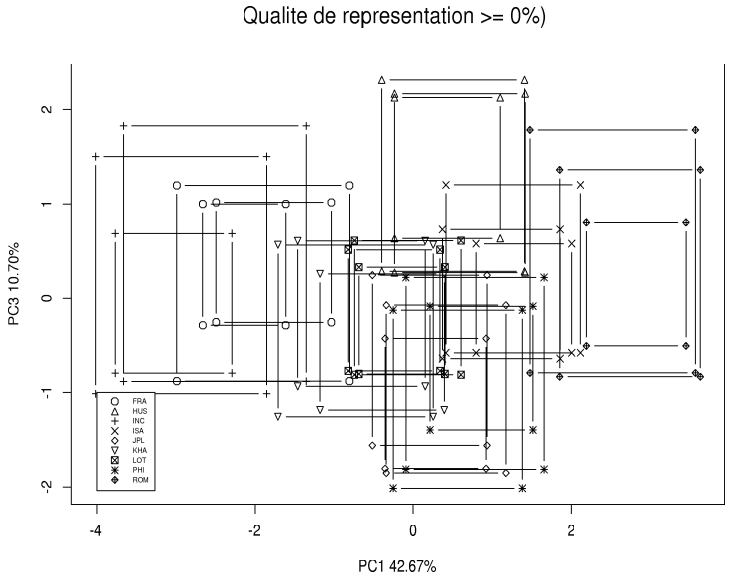


FIG. 3.3: *Deuxième plan factoriel de niveau 0*

Une meilleure représentation factorielle

Pour obtenir une représentation des principaux groupements de visages à des niveaux de qualité de représentation plus élevés, on applique la procédure itérative présentée en 1.2.11. On choisit particulièrement les niveaux $\alpha = 0.2\%$ puis $\alpha = 0.6\%$. La représentation des visages avec une qualité supérieure ou égale à 20% dans les deux premiers plans factoriels est donnée par les figures 3.4 et 3.5. De manière similaire, on visualise dans les figures 3.6 et 3.7 les 27 séquences de visages avec une qualité de représentation supérieure ou égale à 60%.

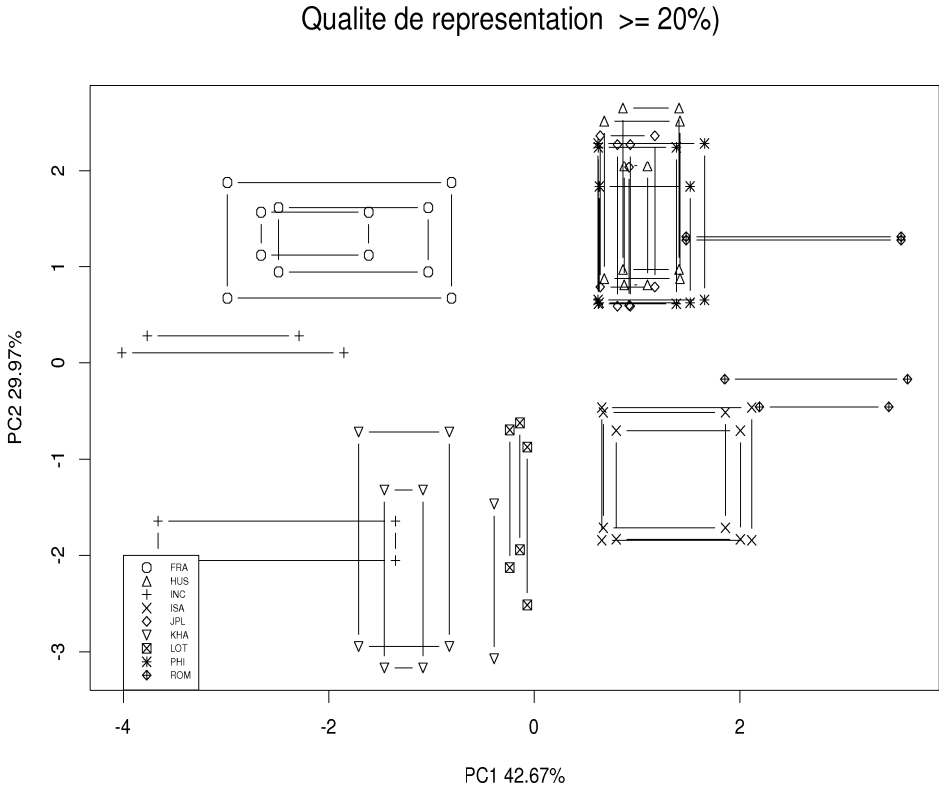


FIG. 3.4: *Premier plan factoriel de niveau 0.2*

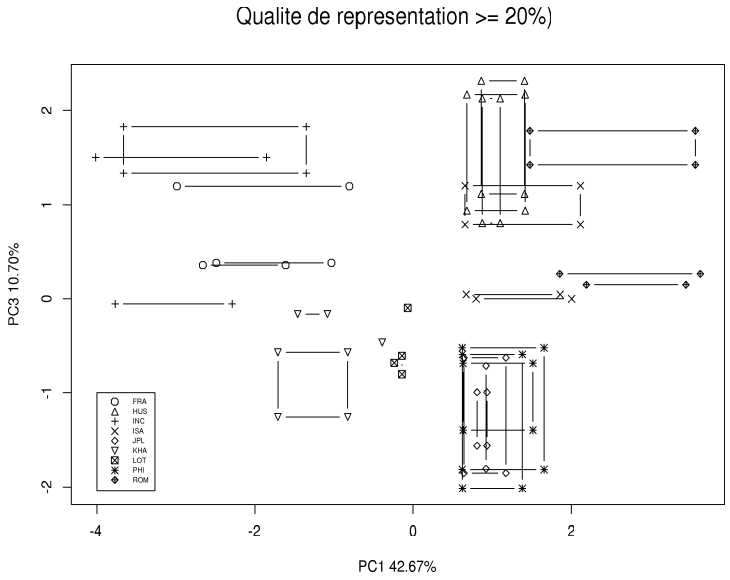


FIG. 3.5: *Deuxième plan factoriel de niveau 0.2*

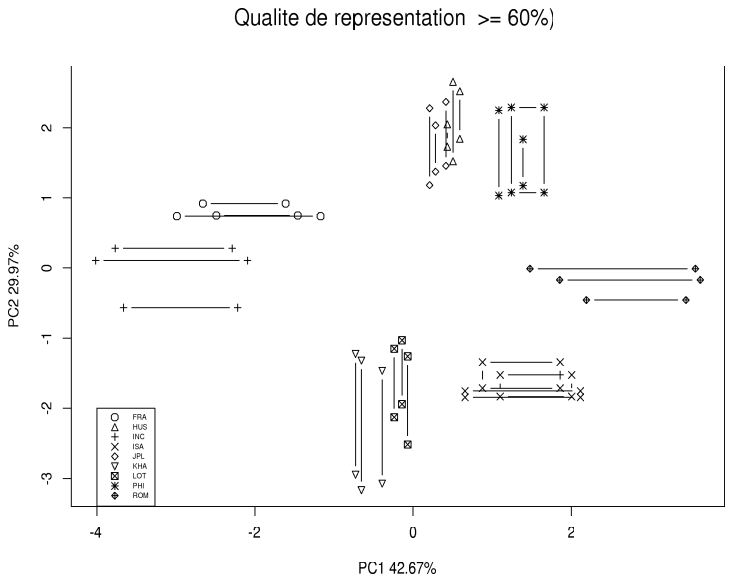


FIG. 3.6: *Premier plan factoriel de niveau 0.6*

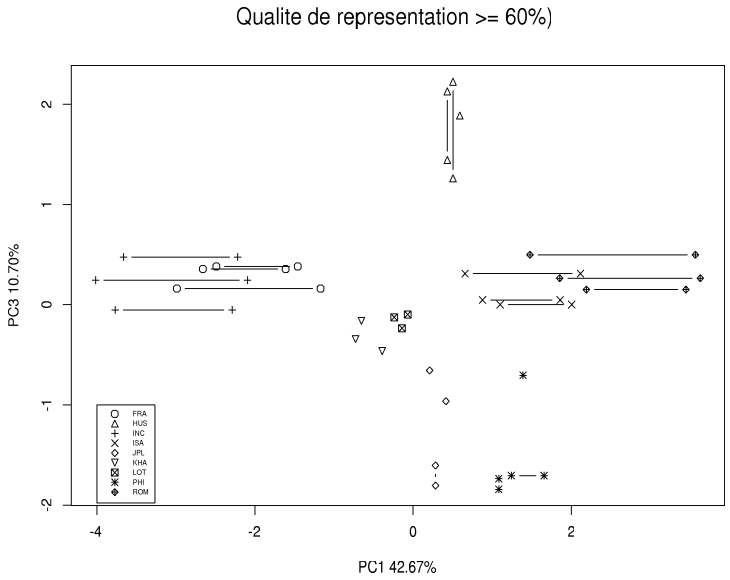


FIG. 3.7: *Deuxième plan factoriel de niveau 0.6*

Corrélations variables/composantes principales

L'étude des corrélations (voir tableau 3.4) nous révèle une forte liaison, d'une part, entre la première composante principale CP_1 et les variables DH à 89%, AH à 83% et AD à 64% et d'autre part, entre la deuxième composante principale CP_2 et les variables, GH à 75%, BC à 66% et EH à 62%.

	Méthode des sommets			
	CP1	CP2	CP3	CP4
AD	0.6444	0.5889	0.1717	-0.1771
BC	0.4903	0.6663	-0.1403	0.5375
AH	0.8374	-0.1968	-0.3707	-0.1884
DH	0.8913	0.0885	0.1649	-0.1954
EH	-0.4749	0.6248	-0.5607	-0.2268
GH	-0.4283	0.7554	0.3377	-0.1746

TAB. 3.4: *Corrélations entre les variables descriptives et les composantes principales*

Paramètres d'aide à l'interprétation

Les contributions relatives, absolues et à l'inertie totale figurent, respectivement, dans les tableaux ci-dessous. Rappelons que les paramètres d'aide à l'interprétation de chaque objet sont définis à partir des paramètres d'aide à l'interprétation de l'ensemble des sommets de cet objet.

	Méthode des sommets Contributions relatives			
	CP1	CP2	CP3	CP4
FRA1	**0.70**	0.14	0.04	0.02
FRA2	**0.62**	0.15	0.04	0.08
FRA3	**0.64**	0.14	0.05	0.07
HUS1	0.06	**0.33**	**0.41**	0.07
HUS2	0.07	**0.45**	**0.32**	0.05
HUS3	0.10	**0.40**	**0.29**	0.03
INC1	**0.86**	0.03	0.01	0.04
INC2	**0.61**	0.08	0.08	0.10
INC3	**0.77**	0.04	0.05	0.05
ISA1	**0.51**	**0.33**	0.02	0.07
ISA2	**0.42**	**0.28**	0.03	0.17
ISA3	**0.44**	**0.27**	0.07	0.12
JPL1	0.04	**0.42**	**0.33**	0.06
JPL2	0.08	**0.43**	**0.21**	0.11
JPL3	0.05	**0.50**	**0.14**	0.16
KHA1	0.04	**0.78**	0.05	0.09
KHA2	0.08	**0.77**	0.03	0.08
KHA3	0.12	**0.60**	0.06	0.14
LOT1	0.02	**0.68**	0.04	0.12
LOT2	0.02	**0.54**	0.05	0.20
LOT3	0.03	**0.53**	0.05	0.26
PHI1	**0.24**	**0.41**	0.16	0.09
PHI2	**0.21**	**0.43**	0.19	0.07
PHI3	0.14	**0.37**	**0.29**	0.09
ROM1	**0.86**	0.04	0.01	0.07
ROM2	**0.83**	0.04	0.05	0.05
ROM3	**0.73**	0.06	0.08	0.05

	Méthode des sommets Contributions absolues			
	CP1	CP2	CP3	CP4
FRA1	0.07	0.02	0.01	0.01
FRA2	0.05	0.02	0.01	0.03
FRA3	0.06	0.02	0.01	0.03
HUS1	0.00	0.03	0.12	0.02
HUS2	0.01	0.06	0.11	0.02
HUS3	0.01	0.05	0.10	0.01
INC1	0.13	0.01	0.01	0.03
INC2	0.09	0.02	0.05	0.07
INC3	0.13	0.01	0.03	0.04
ISA1	0.03	0.03	0.00	0.02
ISA2	0.02	0.02	0.01	0.05
ISA3	0.02	0.02	0.02	0.03
JPL1	0.00	0.04	0.08	0.02
JPL2	0.00	0.05	0.07	0.04
JPL3	0.00	0.04	0.04	0.05
KHA1	0.00	0.11	0.02	0.05
KHA2	0.01	0.11	0.01	0.03
KHA3	0.01	0.08	0.02	0.05
LOT1	0.00	0.06	0.01	0.04
LOT2	0.00	0.04	0.01	0.05
LOT3	0.00	0.04	0.01	0.09
PHI1	0.01	0.03	0.04	0.02
PHI2	0.01	0.04	0.05	0.02
PHI3	0.01	0.04	0.08	0.03
ROM1	0.12	0.01	0.01	0.06
ROM2	0.11	0.01	0.03	0.04
ROM3	0.09	0.01	0.04	0.04

Discussion

Dans le premier plan factoriel de niveau 0 (figure 3.2), on distingue quatre ou cinq classes de visages : {Rom, Isa}, {Phi, Jpl, Hus}, {Fra, Inc} {Lot, Kha}. On constate un fort recouvrement entre les trois séquences d'images d'une même personne ce qui traduit une cohérence des descriptions issues des trois bandes-vidéo. La représentation par des rectangles révèle une variation similaire à l'intérieur des classes à l'exception de la classe Inc qui présente une variation légérement plus important que celle des autres classes de visages.

Dans le premier plan factoriel de niveau 0.2 (figure 3.4), les quatre classes de visages sont beaucoup plus apparentes. D'une part, on constate que le groupe {Phi, Jpl, Hus} reste très compacte, ce qui traduit une forte proximité entre les visages Phi, Jpl et Hus avec un degré de confiance de 20%. D'autre part, on constate que Rom et Isa restent également proches à l'exception d'une

	Méthode des sommets Contributions à l'inertie
FRA1	0.04
FRA2	0.03
FRA3	0.04
HUS1	0.03
HUS2	0.04
HUS3	0.04
INC1	0.07
INC2	0.06
INC3	0.07
ISA1	0.03
ISA2	0.02
ISA3	0.02
JPL1	0.03
JPL2	0.03
JPL3	0.03
KHA1	0.04
KHA2	0.04
KHA3	0.04
LOT1	0.03
LOT2	0.02
LOT3	0.02
PHI1	0.02
PHI2	0.03
PHI3	0.03
ROM1	0.06
ROM2	0.06
ROM3	0.06

séquence d'images associée à ROM dont la description se situe proche de celle des visages du groupe {PHI, JPL, HUS}. Les visages LOT et KHA} restent également proche, de même pour les visages FRA et INC. À un niveau 0.6 (figure 3.6), les quatre groupes sont bien mis en évidence, on voit nettement la grande proximité entre les trois séquences d'images d'une même personne.

Le sens des allongements des rectangles, des segments (vertical, horizontal) constitue un élément important pour l'interprétation de la variance intrinsèque à chaque visage. En effet, considérant la figure 3.6, les visages ROM, FRA, INC et ISA, représentés par des segments horizontaux, présentent une grande variation pour les variables fortement corrélées avec le premier axe factoriel CP_1, à savoir DH, AH et AD. Par ailleurs, les visages HUS, KHA et LOT, représentés par des segments verticaux, présentent une grande variation pour les variables GH, BC et EH fortement corrélées avec le deuxième axe factoriel CP_2.

En nous ramenant au sens géométrique des descripteurs du visage, les variables fortement liées à la première composante principale mesurent l'allongement vertical du visage, alors que les variables caractérisant le deuxième axe principal décrivent plutôt la largeur du visage.

D'après ces résulats, les visages allongés ou ovales se projettent dans des régions de fortes valeurs pour la première composante principale, alors que les visages plutôt ronds ou larges se projettent dans des régions de fortes valeurs pour la deuxième composante PC_2 (figure 3.8).

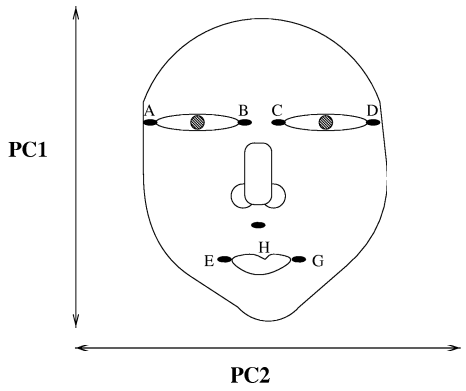


FIG. 3.8:

3.1.4 Résultats dans le cadre de la méthode des centres

L'application de la *méthode des centres* au tableau de données défini en 3.1 revient à effectuer une analyse en composantes principales sur le tableau des centres des intervalles ; les sommets de chaque hyper-rectangle sont projetés a posteriori en supplémentaires. Les valeurs propres ainsi que les pourcentages d'inertie figurent dans le tableau 3.5 suivant.

	Méthode des centres		
Numéro	Valeurs propres	% d'inertie	Cumul
1	2.788	46.5	46.5
2	2.044	34.1	80.6
3	0.547	9.1	89.7
4	0.324	5.4	95.1
5	0.234	3.9	99.9
6	0.062	1.0	100

TAB. 3.5: *Valeurs propres et pourcentage d'inertie*

Description et représentation factorielle des visages

Rappelons que dans la méthode des centres, les composantes principales de type intervalle sont définies à partir des coordonnées factorielles des sommets, éléments supplémentaires dans l'analyse. La description des 27 séquences de visages dans l'espace factoriel des quatre premières composantes principales est donnée par le tableau 3.1.4.

Le principe de la représentation des objets est identique à celui de la méthode des sommets. Les figures 3.9, 3.10 donnent la représentation factorielle des 27 séquences de visages dans, respectivement, les plans factoriels (PC_1, PC_2) et (PC_1, PC_3).

	Méthode des centres			
	CP1	CP2	CP3	CP4
FRA1	[-2.969,-1.747]	[0.236, 1.722]	[-0.376, 1.033]	[-0.778, 0.566]
FRA2	[-2.729,-1.111]	[-0.197, 1.789]	[-0.372, 1.067]	[-0.804, 1.198]
FRA3	[-3.286,-0.856]	[-0.523, 2.077]	[-1.032, 1.301]	[-1.092, 1.414]
HUS1	[-0.374, 1.162]	[0.376, 2.284]	[0.772, 2.419]	[-1.231, 0.538]
HUS2	[-0.583, 1.482]	[0.592, 2.975]	[0.355, 2.592]	[-1.343, 0.757]
HUS3	[-0.403, 1.506]	[0.461, 2.822]	[0.384, 2.418]	[-1.085, 0.680]
INC1	[-4.065,-2.381]	[-0.902, 1.289]	[-0.905, 0.744]	[-0.758, 1.362]
INC2	[-3.909,-1.213]	[-2.509, 0.933]	[-0.954, 2.022]	[-0.986, 2.239]
INC3	[-4.332,-1.837]	[-1.495, 1.469]	[-1.115, 1.615]	[-0.675, 1.786]
ISA1	[0.881, 2.239]	[-2.063,-0.477]	[-0.603, 0.708]	[-1.136, 0.465]
ISA2	[0.388, 2.038]	[-1.933,-0.073]	[-0.688, 0.868]	[-1.556, 0.200]
ISA3	[0.444, 2.355]	[-2.088,-0.112]	[-0.618, 1.388]	[-1.438, 0.455]
JPL1	[-0.567, 0.888]	[0.667, 2.379]	[-2.068,-0.536]	[-1.124, 0.457]
JPL2	[-0.593, 1.167]	[0.575, 2.764]	[-2.101,-0.145]	[-1.523, 0.461]
JPL3	[-0.782, 0.916]	[0.593, 2.645]	[-1.805, 0.206]	[-1.650, 0.441]
KHA1	[-1.097, 0.641]	[-3.507,-1.653]	[-1.280, 0.368]	[-0.212, 1.617]
KHA2	[-1.429, 0.386]	[-3.645,-1.505]	[-0.993, 0.743]	[-0.626, 1.446]
KHA3	[-1.730, 0.468]	[-3.393,-0.817]	[-1.346, 0.714]	[-0.837, 1.743]
LOT1	[-0.794, 0.742]	[-2.864,-0.977]	[-0.874, 0.707]	[-1.647, 0.093]
LOT2	[-0.773, 0.472]	[-2.205,-0.687]	[-0.879, 0.376]	[-1.674,-0.225]
LOT3	[-0.928, 0.408]	[-2.435,-0.778]	[-0.848, 0.586]	[-2.006,-0.390]
PHI1	[0.114, 1.574]	[0.722, 2.178]	[-1.582,-0.030]	[-0.168, 1.156]
PHI2	[-0.270, 1.740]	[0.454, 2.689]	[-2.017, 0.220]	[-0.634, 1.317]
PHI3	[-0.450, 1.440]	[0.356, 2.671]	[-2.258,-0.168]	[-0.677, 1.358]
ROM1	[2.407, 3.838]	[-1.270, 0.436]	[-0.549, 0.902]	[0.198, 1.766]
ROM2	[1.961, 4.041]	[-1.394, 1.200]	[-0.899, 1.491]	[-0.460, 1.811]
ROM3	[1.529, 3.978]	[-1.436, 1.585]	[-0.874, 1.918]	[-0.689, 1.901]

TAB. 3.6: *Les quatres premières composantes principales de type intervalle*

Remarquons que, dans la méthode des centres, on peut également utiliser la procédure itérative puisque les cosinus des sommets, éléments inactifs dans l'analyse, gardent un sens.

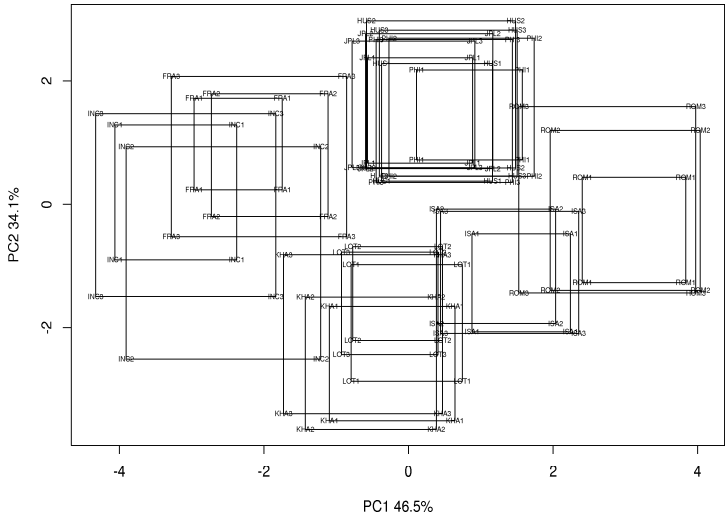


FIG. 3.9:

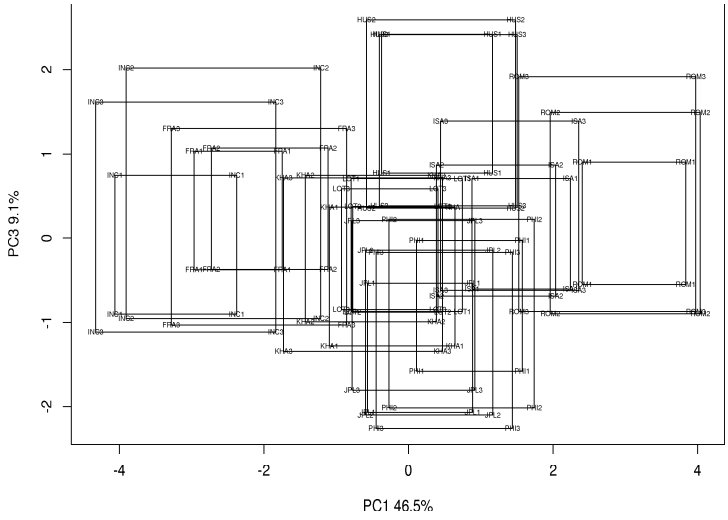


FIG. 3.10:

Corrélations variables/composantes principales

Les corrélations entre les variables et les quatres premières composantes principales sont :

	Méthode des centres			
	CP1	CP2	CP3	CP4
AD	0.640	0.648	0.174	- 0.094
BC	0.496	0.736	- 0.140	0.423
AH	0.862	- 0.164	- 0.418	- 0.189
DH	0.910	0.130	0.205	- 0.193
EH	- 0.559	0.655	- 0.464	- 0.177
GH	- 0.500	0.780	0.255	- 0.180

TAB. 3.7: *Corrélations entre variables descriptives et composantes principales*

Les paramètres d'aide à l'interprétation

Les tableaux ci-dessous donnent les contributions relatives, absolues et à l'inertie totale des 27 séquences de visages projetées. Signalons que, dans la méthode des centres, on estime les paramètres d'aide à l'interprétation des objets par ceux de leur centre.

	Méthode des centres Contributions relatives			
	CP1	CP2	CP3	CP4
FRA1	0.743	0.128	0.014	0.001
FRA2	0.709	0.122	0.023	0.007
FRA3	0.827	0.116	0.004	0.005
HUS1	0.033	0.372	0.535	0.025
HUS2	0.035	0.559	0.382	0.015
HUS3	0.059	0.526	0.383	0.008
INC1	0.953	0.003	0.001	0.008
INC2	0.750	0.071	0.033	0.045
INC3	0.935	0.000	0.006	0.030
ISA1	0.552	0.365	0.001	0.026
ISA2	0.457	0.312	0.003	0.143
ISA3	0.522	0.322	0.039	0.064
JPL1	0.006	0.515	0.377	0.025
JPL2	0.017	0.564	0.255	0.057
JPL3	0.001	0.672	0.164	0.094
KHA1	0.007	0.889	0.028	0.066
KHA2	0.038	0.929	0.002	0.024
KHA3	0.075	0.839	0.019	0.039
LOT1	0.000	0.807	0.002	0.132
LOT2	0.007	0.622	0.019	0.268
LOT3	0.016	0.601	0.004	0.334
PHI1	0.177	0.522	0.161	0.061
PHI2	0.129	0.591	0.193	0.028
PHI3	0.055	0.510	0.328	0.026
ROM1	0.888	0.016	0.003	0.088
ROM2	0.935	0.001	0.009	0.047
ROM3	0.884	0.001	0.032	0.043

	Méthode des centres Contributions absolues			
FRA1	0.074	0.017	0.007	0.001
FRA2	0.049	0.011	0.008	0.004
FRA3	0.057	0.011	0.001	0.003
HUS1	0.002	0.032	0.172	0.014
HUS2	0.003	0.058	0.147	0.010
HUS3	0.004	0.049	0.133	0.005
INC1	0.138	0.001	0.000	0.010
INC2	0.087	0.011	0.019	0.045
INC3	0.126	0.000	0.004	0.035
ISA1	0.032	0.029	0.000	0.013
ISA2	0.020	0.018	0.001	0.052
ISA3	0.026	0.022	0.010	0.028
JPL1	0.000	0.042	0.115	0.013
JPL2	0.001	0.051	0.085	0.032
JPL3	0.000	0.047	0.043	0.042
KHA1	0.001	0.121	0.014	0.056
KHA2	0.004	0.120	0.001	0.019
KHA3	0.005	0.080	0.007	0.023
LOT1	0.000	0.067	0.000	0.069
LOT2	0.000	0.038	0.004	0.103
LOT3	0.001	0.047	0.001	0.164
PHI1	0.009	0.038	0.044	0.028
PHI2	0.007	0.045	0.055	0.013
PHI3	0.003	0.042	0.100	0.013
ROM1	0.129	0.003	0.002	0.110
ROM2	0.120	0.000	0.006	0.052
ROM3	0.101	0.000	0.018	0.042

	Méthode des centres Contributions à l'inertie
FRA1	0.046
FRA2	0.032
FRA3	0.032
HUS1	0.029
HUS2	0.035
HUS3	0.032
INC1	0.067
INC2	0.054
INC3	0.063
ISA1	0.027
ISA2	0.020
ISA3	0.023
JPL1	0.028
JPL2	0.031
JPL3	0.024
KHA1	0.046
KHA2	0.044
KHA3	0.033
LOT1	0.028
LOT2	0.021
LOT3	0.027
PHI1	0.025
PHI2	0.026
PHI3	0.028
ROM1	0.068
ROM2	0.059
ROM3	0.053

Discussion

Rappelons que dans la méthode des centres les axes factoriels d'inertie ne sont déterminés qu'à partir des centres des hyper-rectangles, la variation est restituée (visualisée) en sortie par projection des sommets en tant qu'éléments supplémentaires.

La représentation factorielle des 27 séquences d'images révèle des résultats comparables à ceux obtenus par la méthode des sommets. En effet, on constate d'une part un fort recouvrement des 3 séquences d'images associées à une même personne ; d'autre part, on distingue quatre classes de visages : {Inc, Fra}, {Lot, Kha}, {Rom, Isa} et {Hus, Jpl, Phi}.

Les corrélations rejoignent celles obtenues dans le cas de la méthode des sommets : les variables fortement liées à la première composante principale mesurent l'allongement vertical du visage, alors que les variables caractérisant le deuxième axe principal décrivent plutôt la largeur du visage. Toutes les interprétation faites dans la méthode des sommets restent ici valables.

Remarquons qu'il n'est pas étonnant que les résultats issus de la méthode des sommets soient similaires à ceux obtenus dans le cas de la méthode des centres. En effet, le fait que les visages soient décrits par des intervalles de faibles amplitudes et que les variations intrinsèques aux visages soient presques équivalentes engendrent un nuage des sommets des hyper-rectangles très proche de celui des centres.

3.2 Application en statistique officielle sur la confidentialité des données

3.2.1 Introduction

Il existe des données qui, pour des raisons de sécurité, ne peuvent être divulguées telles qu'elles sont à l'origine. De telles données subissent d'abord un codage avant d'être communiquées en vue de traitements divers.

Le codage des données confidentielles doit masquer les données d'origine tout en conservant au mieux l'information (structure des données, propriétés statistiques des données, etc.) apportée par ces données. Dans le cas où les données sont destinées à être traitées par des méthodes d'analyse des données ou de statistique, l'information apportée s'exprime en terme d'inertie ou de variance. Les techniques de codage des données confidentielles, souvent fondées sur la suppressions des données, l'agrégation et l'estimation par des valeurs centrales etc., sont confrontées au problème de la réduction de la variance.

On présente, dans un premier temps, une technique de codage des données confidentielles fondée sur la *micro-agrégation* des données [Defays et al.93], [Anwar93]. On propose la modification de cette technique par l'utilisation d'un codage par sommets. On compare les variances prises en compte dans le cas de ces deux techniques et on montre le gain effectué sur la variance grâce à l'utilisation du codage par sommets.

3.2.2 Masquage des données par *micro-agrégation*

Le principe du codage par micro-agrégation se décompose en trois étapes. Dans la première, on trie les observations prises par une variable donnée et on étiquette chaque individu du rang de l'observation associée. Dans la deuxième, on construit des groupes de k observations successives, à partir des observations triées. Finalement, dans la dernière étape, on code chaque observation par la moyenne des observations du groupe auquel elle appartient. Dans le cas de plusieurs variables, on procède au même traitement séparément pour chacune.

Soit $x_1, \ldots, x_n$ les valeurs prises par la variable X pour n individus. On suppose, ce qui ne restreint pas la généralité, que les observations sont triées. L'agrégation des k observations successives et le codage de chaque groupe par la valeur la plus représentative (i.e. le centre, le mode, etc.) sont donnés par le tableau suivant :

N^o groupe	X	Poids	Centre	Codage
1	x_1^1	p_1^1		x_c^1
	$\vdots$	$\vdots$	x_1^c	$\vdots$
	$x_{n_1}^1$	$p_{n_1}^1$		x_c^1
$\vdots$	$\ddots$	$\vdots$		
r	x_1^r	p_1^r		x_c^r
	$\vdots$	$\vdots$	x_r^c	$\vdots$
	$x_{n_r}^r$	$p_{n_r}^r$		x_c^r

TAB. 3.8: *Données confidentielles d'origine et codage par les centres des groupes*

On note $p_j = \sum_{i=1}^{n_j} p_i^j$ le poids du groupe j.

3.2.3 Micro-agrégation par intervalle

L'idée est simple : après la construction des groupes en k observations, on propose de coder chaque observation par l'intervalle de variation incluant toutes les observations du groupe auquel elle appartient. On procède de même pour l'ensemble des variables quantitatives. Un codage par sommets est ensuite effectué.

N^o groupe	X	Poids	Intervalle	Codage
1	x_1^1	p_1^1	$[\underline{x_1}, \overline{x_1}]$	$[\underline{x_1}, \overline{x_1}]$
	$\vdots$	$\vdots$		$\vdots$
	$x_{n_1}^1$	$p_{n_1}^1$		$[\underline{x_1}, \overline{x_1}]$
$\vdots$	$\ddots$	$\vdots$		
r	x_1^r	p_1^r	$[\underline{x_r}, \overline{x_r}]$	$[\underline{x_r}, \overline{x_r}]$
	$\vdots$	$\vdots$		$\vdots$
	$x_{n_r}^r$	$p_{n_r}^r$		$[\underline{x_r}, \overline{x_r}]$

TAB. 3.9: *Données confidentielles d'origine et codage par intervalles*

3.2.4 Comparaisons des variances

On s'intéresse dans cette section à la comparaison de la variance prise en compte à la suite d'un codage par les centres et celle issue d'un codage fondé sur les intervalles de variation caractérisant les groupes. Pour tenir compte de la distribution intrinsèque à chaque groupe, on propose de pondérer les bornes de chaque intervalle $[\underline{x_j}, \overline{x_j}]$ par, respectivement, les coefficients $\underline{p_j}$ et $\overline{p_j}$:

$$\begin{aligned} \underline{p_j} + \overline{p_j} &= 1 \\ \underline{p_j x_j} + \overline{p_j x_j} &= x_c^j \end{aligned}$$

ainsi, le centre de gravité des bornes $\underline{x_j}$ et $\overline{x_j}$ coïncide avec la valeur représentant au mieux cet intervalle. Dans le cas d'un codage par les centres, la variance V_c prise en compte est :

$$V_c = \sum_{j=1}^{r} \sum_{i=1}^{n_j} p_i^j (x_c^j - \overline{X_c})^2 \tag{3.1}$$

$$= \sum_{j=1}^{r} (x_c^j - \overline{X_c})^2 (\sum_{i=1}^{n_j} p_j^j) = \sum_{j=1}^{r} (x_c^j - \overline{X_c})^2 p^j \tag{3.2}$$

où $\overline{X_c}$ est la moyenne à la suite du codage par les centres, p_i^j le poids de l'individu i du groupe j et p_j le poids du groupe j. D'autre part, la variance prise en compte à la suite d'un codage par sommets est :

$$V_s = \sum_{j=1}^{r} \sum_{i=1}^{n_j} (p_i^j \underline{p_j} (\underline{x_j} - \overline{X_s})^2 + p_i^j \overline{p_j} (\overline{x_j} - \overline{X_s})^2) \tag{3.3}$$

$$= \sum_{j=1}^{r} p^j (\underline{p_j} (\underline{x_j} - \overline{X_s})^2 + \overline{p_j} (\overline{x_j} - \overline{X_s})^2) \tag{3.4}$$

où $\overline{X_s}$ est la moyenne issue d'un codage par sommets. Nous vérifions aisément que les moyennes $\overline{X_s}$, $\overline{X_c}$ sont égales à la moyenne $\overline{X}$ des données confidentielles avant le codage. L'écart entre les variances V_s et V_c est alors :

$$V_s - V_c = \sum_{j=1}^{r} p^j \underline{p_j} \overline{p_j} (\overline{x_j} - \underline{x_j})^2 \tag{3.5}$$

ou encore :

$$V_s \quad = \quad V_c + E \tag{3.6}$$

où $E = \sum_{j=1}^{r} p^j \underline{p_j} \overline{p_j} (\overline{x_j} - \underline{x_j})^2$ désigne le facteur amplitude.

La démonstration de la relation établie en 3.5 est la même que celle effectuée dans la section 1.5 pour comparer la variance dans la méthode des sommets et des centres. Comme $E \geq 0$, alors :

$$V_s \geq V_c \tag{3.7}$$

Ce résultat montre que la variance V_c issue d'un codage des données confidentielles par les centres des groupes est toujours inférieure à la variance V_s issue d'un codage des groupes par l'intervalle décrivant la variation au sein du groupe. Cette proposition constitue une ouverture quant à l'utilisation des données intervalles pour le masquage des données confidentielles.

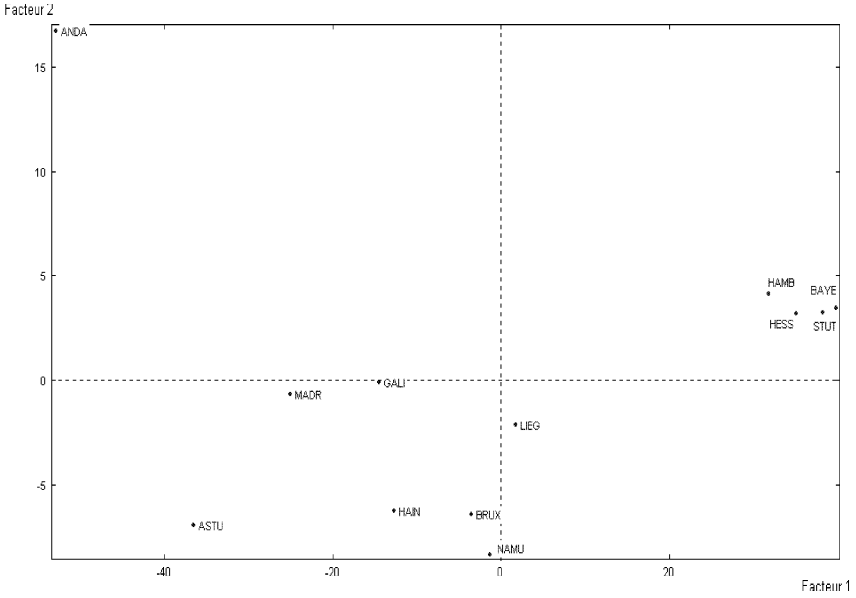
Ce travail est en cours de développement et d'important aspects restent à explorer et à étudier pour l'enrichissement et l'optimisation des méthodes de codages des données confidentielles. Dans le cas, par exemple, où les valeurs extrêmes prises par les groupes ne peuvent pas également être divulguées, on pense à un type de normalisation des variables intervalles qui puisse, tout en masquant les valeurs extrêmes, conserver les propriétés statistiques de ces variables.

3.2.5 Application à des statistiques européennes sur le chômage

Considérons le tableau suivant donnant la description du chomâge des hommes, des femmes, des moins de 25 ans et des plus de 25 ans dans les régions de trois pays européens en avril 1995 et en avril 1996. Ces données sont extraites de la publication "Statistiques en bref" parue en 1996 [Eurostat97] :

Pays	Régions	Homme		Femme		< 25 ans		> 25ans	
		95	96	95	96	95	96	95	96
Belgique	BRUX	12.8	13.6	14.0	14.8	33.3	34.4	11.4	12.5
	HAIN	13.3	13.1	19.6	19.5	39.2	37.8	12.9	13.2
	LIEG	12.7	10.3	15.9	16.6	29.4	28.1	10.7	11.2
	NAMU	10.2	9.6	15.4	15.2	35.4	30.9	9.8	9.9
Allemagne	STUT	5.2	5.7	5.5	5.4	6.2	6.8	5.2	5.4
	BAYE	4.7	5.4	5.2	5.2	4.9	6.2	4.7	5.1
	HAMB	7.7	9.1	6.6	6.8	8.9	10.7	7.0	7.8
	HESS	6.0	6.7	6.3	6.2	7.5	9.1	5.9	6.2
Espagne	GALI	13.9	15.5	21.6	24.0	34.2	39.3	14.7	15.9
	ASTU	17.0	18.0	27.6	29.3	51.5	51.9	16.4	17.6
	MADR	17.2	16.4	26.1	27.0	40.3	45.1	17.1	16.2
	ANDA	28.1	27.2	42.4	41.0	50.9	49.4	28.7	28.4

On applique l'ACP classique au tableau ci-dessus ; la représentation des 12 régions dans le premier plan factoriel est :



Micro-agrégation par les centres

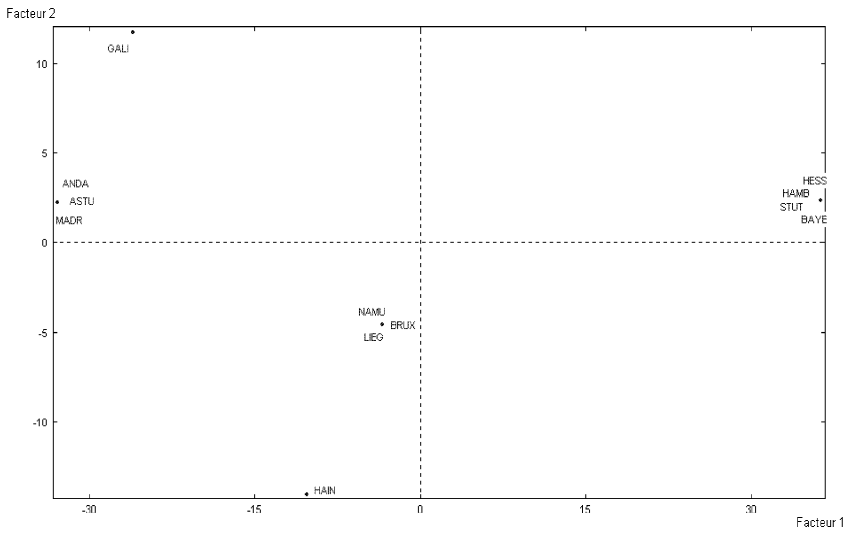
On procède au tri, par ordre croissant, des observations de chaque variable. On construit ensuite des groupes de $k = 4$ observations successives. On constate que chaque groupe ainsi construit se compose des observations des régions d'un même pays, à l'exception des groupes associés à la variable représentant les moins de 25 ans en 1995. La micro-agrégation par les centres revient à substituer chaque observation du tableau des données d'origine par la moyenne des observations du groupe auquel elle appartient. Les données ainsi codées sont :

Régions	Homme		Femme		< 25 ans		> 25 ans	
	95	96	95	96	95	96	95	96
BRUX	12.25	11.65	16.225	16.525	33.075	32.8	11.2	11.7
HAIN	12.25	11.65	16.225	16.525	45.475	32.8	11.2	11.7
LIEG	12.25	11.65	16.225	16.525	33.075	32.8	11.2	11.7
NAMU	12.25	11.65	16.225	16.525	33.075	32.8	11.2	11.7
STUT	5.9	6.725	5.9	5.9	6.875	8.2	5.7	6.125
BAYE	5.9	6.725	5.9	5.9	6.875	8.2	5.7	6.125
HAMB	5.9	6.725	5.9	5.9	6.875	8.2	5.7	6.125
HESS	5.9	6.725	5.9	5.9	6.875	8.2	5.7	6.125
GALI	19.05	19.275	29.425	30.325	33.075	46.425	19.225	19.525
ASTU	19.05	19.275	29.425	30.325	45.475	46.425	19.225	19.525
MADR	19.05	19.275	29.425	30.325	45.475	46.425	19.225	19.525
ANDA	19.05	19.275	29.425	30.325	45.475	46.425	19.225	19.525

La matrice de variance-covariance associée est :

H95	28.83							
H96	27.57	26.66						
F95	51.63	49.64	92.70					
F96	53.61	51.57	96.27	99.99				
M2595	77.23	70.79	135.50	140.44	258.91			
M2596	83.50	78.31	148.12	153.67	240.25	250.22		
P2595	29.71	28.67	53.43	55.50	76.78	84.63	30.84	
P2596	29.42	28.37	52.90	54.95	76.40	84.00	30.52	30.21

où les variables H95, H96, F95, F96, M2595,M2596, P2595, P2596 décrivent le chômage, respectivement, pour les hommes en 95 et en 96, pour les femmes en 95 et en 96, pour les moins de 25 ans en 95 et en 96, finalement pour les plus de 25 ans en 95 et en 96. On applique l'Acp classique au tableau issu de la micro-agrégation par les centres ; la représentation des 12 régions dans le premier plan factoriel est alors :



Micro-agrégation des données confidentielles par intervalles

Après le tri de chaque variable et la construction des groupes composés des $k = 4$ valeurs successsives, la micro-agrégation par intervalles consiste à coder chaque observation du tableau de départ par le plus petit intervalle incluant toutes les observations du groupe auquel elle appartient. Les données ainsi codées sont :

Régions	Homme		Femme		< 25 ans		> 25 ans	
	95	96	95	96	95	96	95	96
BRUX	[10.20,13.30]	[9.60,13.60]	[14.00,19.60]	[14.80,19.50]	[29.40,35.40]	[28.10,37.80]	[9.80,12.90]	[9.90,13.20]
HAIN	[10.20,13.30]	[9.60,13.60]	[14.00,19.60]	[14.80,19.50]	[39.20,51.50]	[28.10,37.80]	[9.80,12.90]	[9.90,13.20]
LIEGE	[10.20,13.30]	[9.60,13.60]	[14.00,19.60]	[14.80,19.50]	[29.40,35.40]	[28.10,37.80]	[9.80,12.90]	[9.90,13.20]
NAMU	[10.20,13.30]	[9.60,13.60]	[14.00,19.60]	[14.80,19.50]	[29.40,35.40]	[28.10,37.80]	[9.80,12.90]	[9.90,13.20]
STUT	[4.70,7.70]	[5.40,9.10]	[5.20,6.60]	[5.20,6.80]	[6.20,8.90]	[6.20,10.70]	[5.20,7.00]	[5.10,7.80]
BAYE	[4.70,7.70]	[5.40,9.10]	[5.20,6.60]	[5.20,6.80]	[6.20,8.90]	[6.20,10.70]	[5.20,7.00]	[5.10,7.80]
HAMB	[4.70,7.70]	[5.40,9.10]	[5.20,6.60]	[5.20,6.80]	[6.20,8.90]	[6.20,10.70]	[5.20,7.00]	[5.10,7.80]
HESS	[4.70,7.70]	[5.40,9.10]	[5.20,6.60]	[5.20,6.80]	[6.20,8.90]	[6.20,10.70]	[5.20,7.00]	[5.10,7.80]
GALI	[13.90,28.10]	[15.50,27.20]	[21.60,42.40]	[24.00,41.00]	[29.40,35.40]	[39.30,51.90]	[14.70,28.70]	[15.90,28.40]
ASTU	[13.90,28.10]	[15.50,27.20]	[21.60,42.40]	[24.00,41.00]	[39.20,51.50]	[39.30,51.90]	[14.70,28.70]	[15.90,28.40]
MADR	[13.90,28.10]	[15.50,27.20]	[21.60,42.40]	[24.00,41.00]	[39.20,51.50]	[39.30,51.90]	[14.70,28.70]	[15.90,28.40]
ANDA	[13.90,28.10]	[15.50,27.20]	[21.60,42.40]	[24.00,41.00]	[39.20,51.50]	[39.30,51.90]	[14.70,28.70]	[15.90,28.40]

La matrice de variance-covariance (des sommets) associée est :

H95	55.62							
H96	35.89	48.64						
F95	5.26	62.63	153.39					
F96	66.23	63.53	116.28	144.16				
M2595	80.81	74.74	145.18	147.61	262.22			
M2596	89.20	83.75	158.77	161.31	228.23	260.58		
P2595	39.53	38.19	69.08	70.09	83.74	93.23	59.41	
P2596	39.86	38.55	69.61	70.62	83.83	93.59	42.38	57.30

On applique la méthode des sommets aux données issues de la micro-agrégation par intervalles ; la figure 3.11 donne la représentation des 12 régions dans le premier plan factoriel.

Discussion

Dans le cas de la micro-agrégation par les centres, la représentation, dans le premier plan factoriel, des 12 régions est assez réductrice ; on a aucune appréciation concernant la variation du chômage au sein de ces pays. Ce plan renvoie l'évolution du chômage entre les trois pays : chômage élevé en Espagne, moyen en Belgique et faible en Allemagne. Le premier plan factoriel représentant les données d'origine révèle une grande variation du chômage entre les régions d'Espagne où il est le plus élevé, cette fluctuation est moins importante en Belgique ; en revanche, en Allemagne le chômage atteint les valeurs les plus basses et une variation très faible entre les régions. On retrouve bien ces informations dans le premier plan factoriel obtenu par l'application de la méthode

des sommets aux données issues de la micro-agrégation par intervalles. En conclusion, la micro-agrégation par intervalles a permis, tout en divulguant les observations du chômage rattachées à chaque région, de restituer et de visualiser la variation du chômage entre les trois pays et entre les régions d'un même pays.

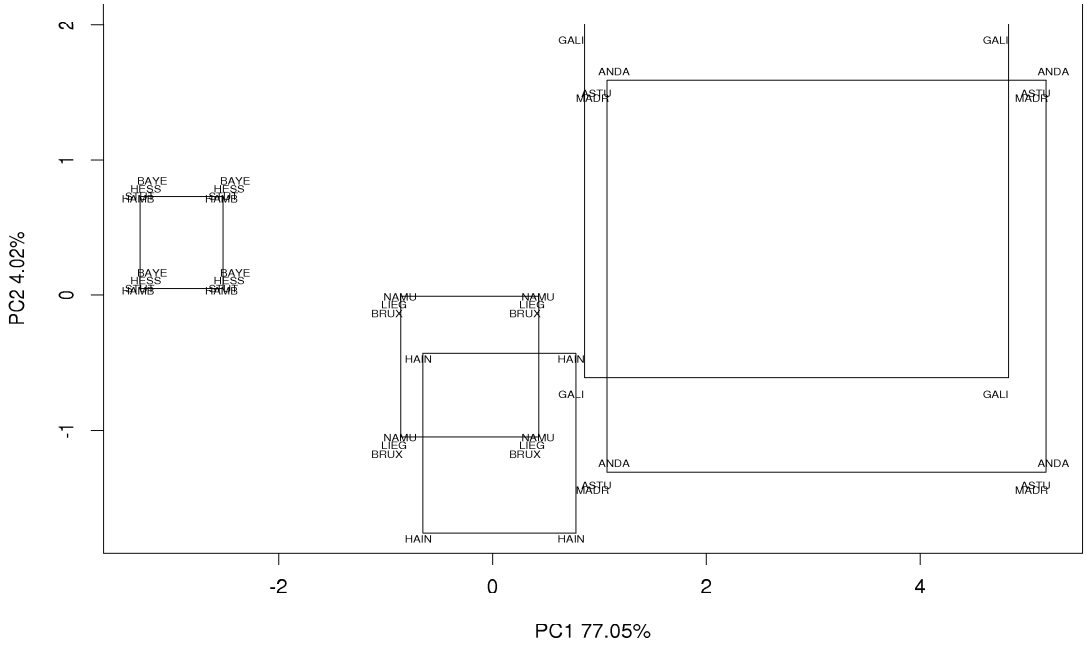


FIG. 3.11:

3.3 Application sur les données *Iris*

L'objet de cette section et de vérifier la validité de la méthode des sommets et de la méthode des centres à travers un jeu de données connu : les *Iris* de FISHER. Les données *Iris* comportent la description de trois classes : *Setosea, Virginica* et *Versicolor*. On dispose de 50 observations pour chaque classe. Les *Iris* sont décrits par quatre variables quantitatives : la longueur et la largeur des pétales, la longueur et la largeur des sépales.

3.3.1 Construction des données intervalles

Pour décrire les *Iris* par des données de type intervalle, on propose de construire, au sein de chaque classe, des groupes de k observations (on choisit $k = 5$) ; chaque groupe ainsi obtenu est décrit par un intervalle défini par la plus petite et la plus grande valeur prise par une variable pour ce groupe. On obtient ainsi un tableau de données à 30 lignes et 4 colonnes (tableau 3.10) donnant la description de 30 objets *Iris* décrits par 4 variables de type intervalle.

Classe	Longueur Sépale	Largeur Sépale	Longueur Pétale	Largeur Pétale
S1	[4.60,5.10]	[3.00,3.60]	[1.30,1.50]	[0.20,0.20]
S2	[4.40,5.40]	[2.90,3.90]	[1.40,1.70]	[0.10,0.40]
S3	[4.30,5.80]	[3.00,4.00]	[1.10,1.60]	[0.10,0.20]
S4	[5.10,5.70]	[3.50,4.40]	[1.30,1.70]	[0.30,0.40]
S5	[4.60,5.40]	[3.30,3.70]	[1.00,1.90]	[0.20,0.50]
S6	[4.70,5.20]	[3.00,3.50]	[1.40,1.60]	[0.20,0.40]
S7	[4.80,5.50]	[3.10,4.20]	[1.40,1.60]	[0.10,0.40]
S8	[4.40,5.50]	[3.00,3.50]	[1.20,1.50]	[0.10,0.20]
S9	[4.40,5.10]	[2.30,3.80]	[1.30,1.90]	[0.20,0.60]
S10	[4.60,5.30]	[3.00,3.80]	[1.40,1.60]	[0.20,0.30]
Ve1	[5.50,7.00]	[2.30,3.20]	[4.00,4.90]	[1.30,1.50]
Ve2	[4.90,6.60]	[2.40,3.30]	[3.30,4.70]	[1.00,1.60]
Ve3	[5.00,6.10]	[2.00,3.00]	[3.50,4.70]	[1.00,1.50]
Ve4	[5.60,6.70]	[2.20,3.10]	[3.90,4.50]	[1.00,1.50]
Ve5	[5.90,6.40]	[2.50,3.20]	[4.00,4.90]	[1.20,1.80]
Ve6	[5.70,6.80]	[2.60,3.00]	[3.50,5.00]	[1.00,1.70]
Ve7	[5.40,6.00]	[2.40,3.00]	[3.70,5.10]	[1.00,1.60]
Ve8	[5.50,6.70]	[2.30,3.40]	[4.00,4.70]	[1.30,1.60]
Ve9	[5.00,6.10]	[2.30,3.00]	[3.30,4.60]	[1.00,1.40]
Ve10	[5.10,6.20]	[2.50,3.00]	[3.00,4.30]	[1.10,1.30]
Vi1	[5.80,7.10]	[2.70,3.30]	[5.10,6.00]	[1.80,2.50]
Vi2	[4.90,7.60]	[2.50,3.60]	[4.50,6.60]	[1.70,2.50]
Vi3	[5.70,6.80]	[2.50,3.20]	[5.00,5.50]	[1.90,2.40]
Vi4	[6.00,7.70]	[2.20,3.80]	[5.00,6.90]	[1.50,2.30]
Vi5	[5.60,7.70]	[2.70,3.30]	[4.90,6.70]	[1.80,2.30]
Vi6	[6.10,7.20]	[2.80,3.20]	[4.80,6.00]	[1.60,2.10]
Vi7	[6.10,7.90]	[2.60,3.80]	[5.10,6.40]	[1.40,2.20]
Vi8	[6.00,7.70]	[3.00,3.40]	[4.80,6.10]	[1.80,2.40]
Vi9	[5.80,6.90]	[2.70,3.30]	[5.10,5.90]	[1.90,2.50]
Vi10	[5.00,6.70]	[2.50,3.40]	[5.00,5.40]	[1.80,2.30]

TAB. 3.10: *Description des données* Iris *par des données intervalles*

3.3.2 Résultats dans le cadre de la méthode des sommets

On applique la *méthode des sommets* aux objets *Iris* décrits dans le tableau 3.10. Les valeurs propres ainsi que les pourcentages d'inertie figurent dans le tableau 3.11.

	Méthode des sommets		
Numéro	Valeurs propres	% d'inertie	Cumul
1	2.6111	65.28	65.28
2	0.8624	21.56	86.84
3	0.4199	10.49	97.33
4	0.1066	2.67	100

TAB. 3.11: *Valeurs propres et pourcentage d'inertie*

Représentation factorielle des *Iris*

La figure 3.12 fournit la représentation factorielle des 30 objets *Iris* dans le plan factoriel de niveau $\alpha = 0$. Appliquons la procédure itérative de visualisation au seuil $\alpha = 0.2$, puis au seuil $\alpha = 0.6$.

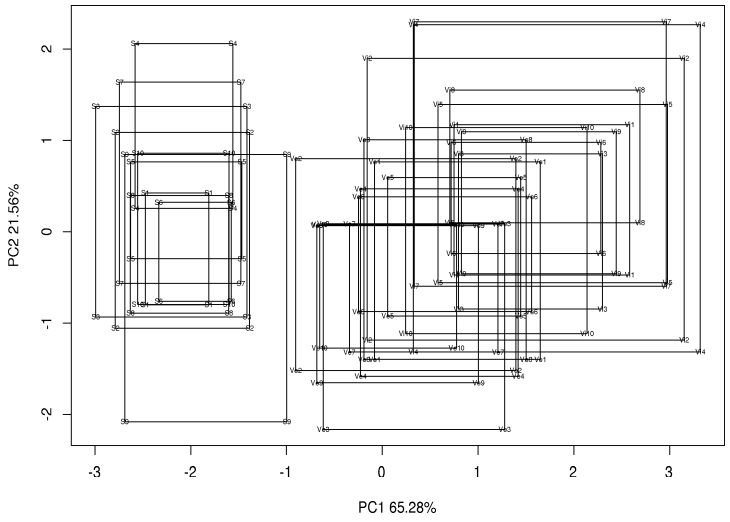


FIG. 3.12: *Représentation des Iris dans le premier plan factoriel de niveau 0*

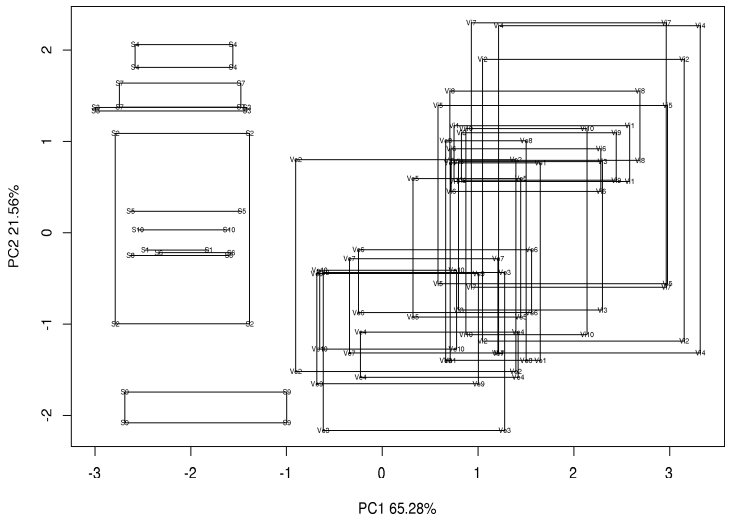


FIG. 3.13: *Représentation des Iris dans le premier plan factoriel de niveau 0.2*

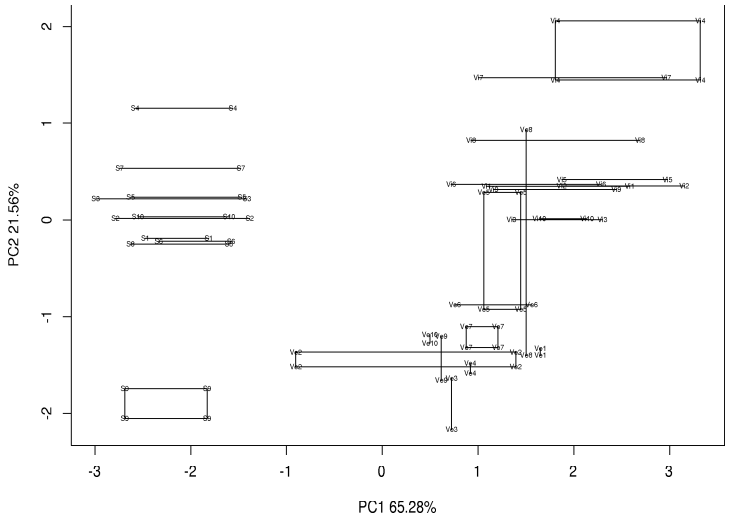


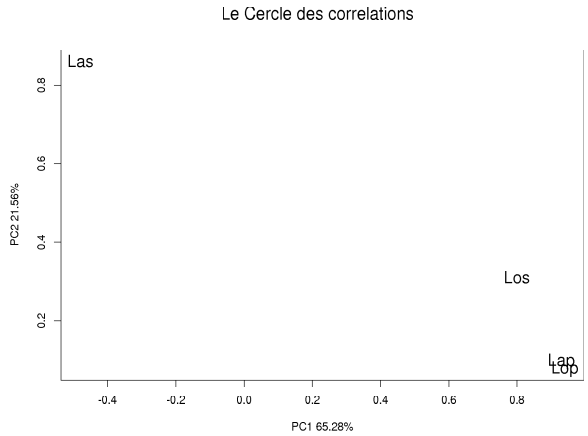
FIG. 3.14: *Représentation des Iris dans le premier plan factoriel de niveau 0.6*

Corrélations variables/composantes principales

L'étude des corrélations montre que le premier axe factoriel caractérise les variables LAP, LOP et LOS ; alors que le deuxième axe caractérise la variable LAS.

	Méthode des sommets			
	CP1	CP2	CP3	CP4
Los	0.7951	0.3130	0.5193	0.0150
Las	-0.4843	0.8646	-0.1334	-0.0100
Lop	0.9409	0.0796	-0.2280	-0.2377
Lap	0.9269	0.1025	-0.2837	0.2232

TAB. 3.12: *Corrélations entre les variables descriptives et les composantes principales*



Discussion

Les plans factoriels de seuils 0, 0.2 et 0.6 révèlent, d'une part, une nette séparation entre la classe *Setosea* et les deux classes *Virginica* et *Versicolor*; d'autre part, un fort recouvrement entre les objets *Iris* issus d'une même classe. Le recouvrement entre les classes *Virginica* et *Versicolor* est fortement présent jusqu'à une qualité de représentation de 60%. La méthode itérative de visualisation permet de distinguer les objets des classes *Virginica* et *Versicolor* qui contribuent le plus au recouvrement entre ces deux classes. On distingue dans le plan de niveau 0.6 les objets Ve_5 et Ve_8 de la classe des *Versicolor* en recouvrement avec les objets Vi_1, Vi_2, Vi_3, Vi_5, Vi_8, Vi_9 et Vi_{10} de la classe

Virginica. D'autre part, le plan factoriel de niveau 0.6 révèle que les objets Ve_8, Ve_5, Ve_3 et Ve_9 de la classe des *Versicolor* présentant un allongement vertical sont caractérisés par une grande variation au niveau de la largeur des sépales (*Las* corrélé à 86% avec PC_2). En revanche, les autres objets présentant un allongement horizontal, particulièrement Ve_2 et S_9 sont caractérisés par une grande variation au niveau de la longueur et la largeur des pétales (*Lap* corrélé à 92% et *Lop* corrélé à 94% avec PC_1).

3.4 Application du codage croisé en ACM

3.4.1 Description des données

On considère la description de 13 races de chiens par 5 variables dont 2 de type intervalle (*Hauteur-garrot*, *Poids*) et 3 variables qualitatives (*Vélocité* à 3 modalités, *Agressivité* à 2 modalités et *Fonctions* à 3 modalités). Les intervalles de variation des variables *Hauteur-garrot* et *Poids* figurent dans les documents de la *Fédération de Cynologie Internationale.*

Races	Hauteur-Garrot	Poids	Vélocité			Agressivité		Fonctions		
			-	+	++	-	+	Compagnie	Chasse	Utile
caniche	[20,35]	[15,25]	0	1	0	1	0	1	0	0
chihuahua	[16,20]	[0.9,3.5]	1	0	0	1	0	1	0	0
pékinois	[20,25]	[3,5]	1	0	0	1	0	1	0	0
basset	[26,40]	[15,23]	1	0	0	0	1	0	1	0
pointer	[60,65]	[25,35]	0	0	1	1	0	0	1	0
setter	[53,62]	[20,32]	0	0	1	1	0	0	1	0
labrador	[54,62]	[20,32]	0	1	0	1	0	0	1	0
lévrier	[55,76]	[25,48]	0	0	1	1	0	0	1	0
mastiff	75	100	1	0	0	0	1	0	0	1
ber-allema	[58,65]	[26,34]	0	0	1	0	1	0	0	1
dog-allema	[76,80]	[50,70]	0	0	1	0	1	0	0	1
doberman	[68,70]	[20,26]	0	0	1	0	1	0	0	1
saint-bern	70	[55 ,80]	1	0	0	0	1	0	0	1

TAB. 3.13: *Description des 13 races de chiens*

3.4.2 Codage linéaire croisé

On note $Hmin$, $Hmax$, $Pmin$,$Pmax$ les variables numériques formées par l'ensemble des bornes inférieures, respectivement, supérieures des variables *Hauteur-Garrot* et *Poids*:

Races	Hmin	Hmax	Pmin	Pmax
caniche	20	35	15	25
chihuahua	16	20	0.9	3.5
pékinois	20	25	3	5
basset	26	40	15	23
pointer	60	65	25	35
setter	53	62	20	32
labrador	54	62	20	32
lévrier	55	76	25	48
mastiff	75	75	100	100
ber-allema	58	65	26	34
dog-allema	76	80	50	70
doberman	68	70	20	26
saint-Bern	70	70	55	80

TAB. 3.14: *Description des 13 races de chiens*

On propose un découpage par histogramme puis l'application d'un codage linéaire des variables numériques obtenues. L'étude des histogrammes (figures 3.15 et 3.16) suggère un découpage des variables $Hmin$,$Hmax$, $Pmin$ et $Pmax$ en 2 modalités, chacune représentant des régions de faibles et de fortes valeurs. La variable $Hmin$, par exemple, est découpée en deux régions : celle des valeurs faibles $[16, 40]$ et celle des valeurs fortes $[40, 76]$.

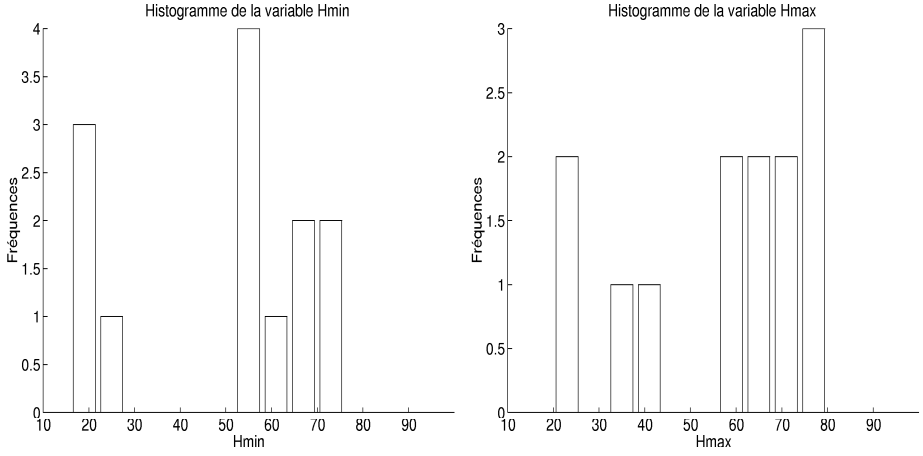


FIG. 3.15: *Histogrammes des variables $Hmin$ et $Hmax$*

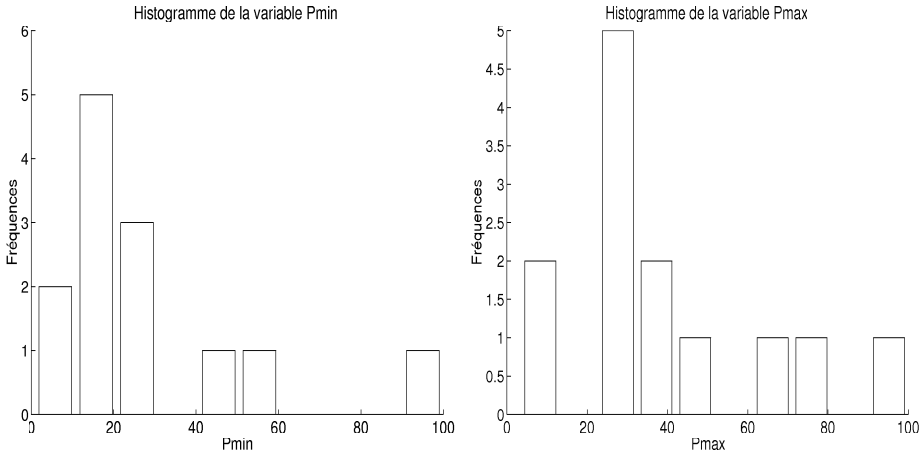


FIG. 3.16: *Histogrammes des variables $Pmin$ et $Pmax$*

Le codage linéaire [Gallego82] en deux modalités des variables numériques $Hmin$, $Hmax$ est alors :

Races	Hmin		Hmax	
	-	+	-	+
caniche	0.9333	0.0667	0.7500	0.2500
chihuahua	1.0000	0	1.0000	0
pékinois	0.9333	0.0667	0.9167	0.0833
basset	0.8333	0.1667	0.6667	0.3333
pointer	0.2667	0.7333	0.2500	0.7500
setter	0.3833	0.6167	0.3000	0.7000
labrador	0.3667	0.6333	0.3000	0.7000
lévrier	0.3500	0.6500	0.0667	0.9333
mastiff	0.0167	0.9833	0.0833	0.9167
ber-allema	0.3000	0.7000	0.250	0 0.7500
dog-allema	0	1.0000	0	1.0000
doberman	0.1333	0.8667	0.1667	0.8333
saint-bern	0.1000	0.9000	0.1667	0.8333

TAB. 3.15: *Codage linéaire en 2 classes : valeurs faibles -, valeurs fortes +*

Le codage linéaire croisé de la variable *Hauteur-Garrot* en (2×2) modalités $H--$, $H-+$, $H+-$ et $H++$ est :

	Hauteur-Garrot			
Races	H- -	H+ -	H - +	H+ +
caniche	0.7000	0.0500	0.2333	0.0167
chihuahua	1.0000	0	0	0
pékinois	0.8556	0.0611	0.0778	0.0056
basset	0.5556	0.1111	0.2778	0.0556
pointer	0.0667	0.1833	0.2000	0.5500
setter	0.1150	0.1850	0.2683	0.4317
labrador	0.1100	0.1900	0.2567	0.4433
lévrier	0.0233	0.0433	0.3267	0.6067
mastiff	0.0014	0.0819	0.0153	0.9014
ber-allema	0.0750	0.1750	0.2250	0.5250
dog-allema	0	0	0	1.0000
doberman	0.0222	0.1444	0.1111	0.7222
saint-bern	0.0167	0.1500	0.0833	0.7500

TAB. 3.16: *Codage linéaire croisé en 4 modalités*

De manière similaire, le codage linéaire des variables $Pmin$, $Pmax$ est :

Races	Pmin		Pmax	
	-	+	-	+
caniche	0.8577	0.1423	0.7772	0.2228
chihuahua	1.0000	0	1.0000	0
pékinois	0.9788	0.0212	0.9845	0.0155
basset	0.8577	0.1423	0.7979	0.2021
pointer	0.7568	0.2432	0.6736	0.3264
setter	0.8073	0.1927	0.7047	0.2953
labrador	0.8073	0.1927	0.7047	0.2953
lévrier	0.7568	0.2432	0.5389	0.4611
mastiff	0	1.0000	0	1.0000
ber-allema	0.7467	0.2533	0.6839	0.3161
dog-allema	0.5045	0.4955	0.3109	0.6891
doberman	0.8073	0.1927	0.7668	0.2332
saint-bern	0.4541	0.5459	0.2073	0.7927

Tab. 3.17: *Codage linéaire en 2 classes : valeurs faibles -, valeurs fortes +*

Le codage linéaire croisé de la variable *Poids* en (2×2) modalités est :

	Poids			
Races	P- -	P+ -	P- +	P+ +
caniche	0.6666	0.1106	0.1911	0.0317
chihuahua	1.0000	0	0	0
pékinois	0.9636	0.0209	0.0152	0.0003
basset	0.6844	0.1135	0.1733	0.0288
pointer	0.5098	0.1638	0.2470	0.0794
setter	0.5689	0.1358	0.2384	0.0569
labrador	0.5689	0.1358	0.2384	0.0569
lévrier	0.4078	0.1310	0.3490	0.1121
mastiff	0	0	0	1.0000
ber-allema	0.5107	0.1732	0.2360	0.0801
dog-allema	0.1569	0.1540	0.3477	0.3414
doberman	0.6190	0.1478	0.1882	0.0449
saint-bern	0.0941	0.1131	0.3600	0.4328

Tab. 3.18: *Codage linéaire croisé en 4 modalités*

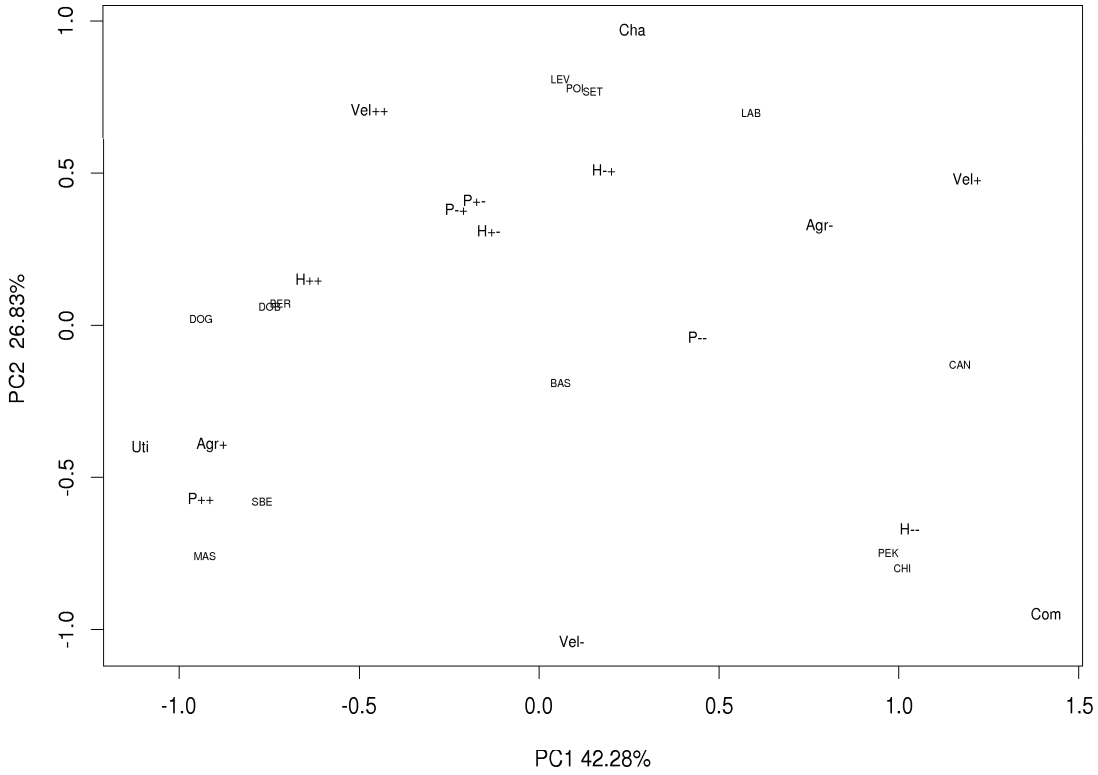
3.4.3 Résultats et Interprétation

Après le codage croisé des deux variables de type intervalle, on applique l'Acm au tableau de données décrit en 3.13 ainsi codé. Les trois premières valeurs propres $\lambda_1 = 0.55$ (42.28%), $\lambda_2 = 0.35$ (26.83%) et $\lambda_3 = 0.17$ (12.78%) restituent 81.89% de l'inertie totale.

Les coordonnées factorielles des chiens ainsi que celles des modalités sur les 3 premiers axes d'inertie figurent dans les tableaux ci-dessous :

	Coordonnées factorielles des races de chiens		
	CP1	CP2	CP3
CAN	1.17	-0.13	0.84
CHI	1.02	-0.80	-0.38
PEK	0.97	-0.74	-0.36
BAS	0.06	-0.20	-0.43
POI	0.10	0.78	-0.35
SET	0.15	0.78	-0.37
LAB	0.59	0.70	0.73
LEV	0.06	0.81	-0.33
MAS	-0.93	-0.76	0.22
BER	-0.72	0.07	0.06
DOG	-0.94	0.02	0.17
DOB	-0.75	0.07	0.06
SBE	-0.77	-0.58	0.13

	Coordonnées factorielles des modalités		
	CP1	CP2	CP3
H–	1.03	-0.67	-0.22
H-+	-0.14	0.31	0.03
H+-	0.17	0.51	-0.01
H++	-0.64	0.15	0.13
P–	0.44	-0.04	-0.16
P-+	-0.19	0.41	0.08
P+-	-0.24	0.38	0.10
P++	-0.94	-0.58	0.33
Vel-	0.09	-1.05	-0.41
Vel+	1.19	0.48	1.92
Vel++	-0.47	0.71	-0.31
Agr-	0.78	0.34	-0.07
Agr+	-0.91	-0.39	0.09
Com	1.42	-0.94	0.09
Cha	0.26	0.97	-0.37
Uti	-1.11	-0.40	0.31



Les paramètres d'aide à l'interprétation

	Contributions absolues des modalités		
	CP1	CP2	CP3
H–	0.105	0.071	0.016
H-+	0.001	0.006	0.000
H+-	0.002	0.024	0.000
H++	0.068	0.006	0.008
P–	0.037	0.000	0.017
P-+	0.001	0.011	0.001
P+-	0.004	0.017	0.002
P++	0.056	0.033	0.024
Vel-	0.001	0.240	0.074
Vel+	0.079	0.020	0.690
Vel++	0.037	0.134	0.054
Agr-	0.120	0.035	0.004
Agr+	0.139	0.040	0.004
Com	0.168	0.120	0.002
Cha	0.010	0.209	0.060
Uti	0.172	0.035	0.045

	Contributions relatives des modalités		
	CP1	CP2	CP3
H–	0.635	0.274	0.029
H-+	0.046	0.239	0.003
H+-	0.066	0.541	0.000
H++	0.771	0.043	0.029
P–	0.631	0.005	0.089
P-+	0.108	0.581	0.022
P+-	0.143	0.392	0.023
P++	0.372	0.137	0.048
Vel-	0.005	0.681	0.099
Vel+	0.257	0.042	0.677
Vel++	0.189	0.433	0.083
Agr-	0.712	0.130	0.006
Agr+	0.712	0.130	0.006
Com	0.600	0.271	0.002
Cha	0.043	0.592	0.081
Uti	0.770	0.101	0.060

	Contributions absolues des races de chiens		
	CP1	CP2	CP3
CAN	0.191	0.004	0.330
CHI	0.144	0.141	0.066
PEK	0.131	0.121	0.059
BAS	0.001	0.009	0.085
POI	0.001	0.135	0.056
SET	0.003	0.133	0.063
LAB	0.049	0.107	0.246
LEV	0.001	0.143	0.049
MAS	0.121	0.129	0.022
BER	0.073	0.001	0.002
DOG	0.124	0.000	0.014
DOB	0.078	0.001	0.002
SBE	0.084	0.075	0.008

	Contributions relatives des races de chiens		
	CP1	CP2	CP3
CAN	0.618	0.008	0.323
CHI	0.549	0.341	0.076
PEK	0.562	0.330	0.076
BAS	0.003	0.038	0.174
POI	0.012	0.775	0.153
SET	0.028	0.763	0.174
LAB	0.211	0.293	0.320
LEV	0.004	0.761	0.124
MAS	0.436	0.294	0.023
BER	0.609	0.006	0.004
DOG	0.785	0.000	0.026
DOB	0.623	0.005	0.004
SBE	0.523	0.299	0.014

Discussion

Le codage linéaire croisé de la hauteur au garrot défini dans le tableau 3.16 montre que la modalité *H-* - caractérise essentiellement *Chi* (1,00), *Pek* (0.85) et *Can* (0.7) traduisant une hauteur au garrot faible et variant très peu chez ces races. La modalité *H-+* caractérise *Lev* (0.32), Bas (0.27) et Set (0.26) et traduit une grande variations de la hauteur au garrot au sein de ces races. Finalement, la modalité *H++* caractérise essentiellement *DogAll* (1,00), *Mas* (0.9) et *SBer* (0.75) traduisant une hauteur au garrot très élevé et variant peu chez ces races.

Le codage linéaire croisé de la variable poids défini dans le tableau 3.18 montre que la modalité *P-* - (variation faible et régions des valeurs basses) caractérise essentiellement *Chi* (1.00), *Pek* (0.96) et *Can* (0.66) traduisant un poids léger et une variation très faible du poids au sein de ces races. D'autre

par, la modalité *P*- - caractérise particulièrement *Mas* (1.00) traduisant un poids très élevé et une variation très faible du poids au sein de cette race.

Par ailleur, on distingue dans le premier plan factoriel quatre principales classes de chiens : {*Mas, Sbe, Dog, Dob, Ber*}, {*Lev, Poi, Set, Lab*}, {*Chi, Pek, Can*} et {*Bas*}. Le premier groupe est caractérisé par un poids élevé (*P*++), agressifs (Agr+), utile (Uti), et dont la hauteur au garrot est élevée (*H*++). Ce groupe peut se scinder en deux sous-groupes : le sous-groupe {*Dog, Ber, Dob*} des chiens véloces, opposés au sous-groupe {*Mas, Sbe*} des chiens plutôt lents. Le second groupe désigne la classe des chiens de chasse (Cha), non agressifs (Agr-), véloces (Vel++), de poids et de hauteur au garrot moyens (*H*-+, *H*+-, *P*-+,*P*+-). Le troisième groupe représente la classe des chiens de compagnie (Com), non agressifs (Agr-), légers (*P*- -) et dont la hauteur au garrot est faible (*H*- -). Le dernier groupe, constitué par les bassets, se différencie du troisième groupe par le caractère agressif (Agr+).

3.5 Application du codage par sommets en ACM

3.5.1 Codage par sommets

Appliquons le codage par sommets défini dans la section 2.3 au tableau de données 3.13. Rappelons tout d'abord qu'un objet (dans ce cas une race de chien) décrit par q observations de type intervalle peut être représenté par un hyper-rectangle à 2^q sommets. La première étape consiste à décomposer chaque ligne du tableau (la description d'un objet) impliquant des données intervalles en un ensemble de lignes, chacune donnant la description d'un des sommets de l'hyper-rectangle associé à l'objet en question. À l'issue de cette première étape, on obtient le tableau suivant :

Races	Hauteur-Garrot	Poids	Vélocité			Agressivité		Fonctions		
			-	+	++	-	+	Compagnie	Chasse	Utile
caniche	20	15	0	1	0	1	0	1	0	0
caniche	35	15	0	1	0	1	0	1	0	0
caniche	20	25	0	1	0	1	0	1	0	0
caniche	35	25	0	1	0	1	0	1	0	0
chihuahua	16	0.9	1	0	0	1	0	1	0	0
chihuahua	20	0.9	1	0	0	1	0	1	0	0
chihuahua	16	3.5	1	0	0	1	0	1	0	0
chihuahua	20	3.5	1	0	0	1	0	1	0	0
pékinois	20	3	1	0	0	1	0	1	0	0
pékinois	25	3	1	0	0	1	0	1	0	0
pékinois	20	5	1	0	0	1	0	1	0	0
pékinois	25	5	1	0	0	1	0	1	0	0
basset	26	15	1	0	0	0	1	0	1	0
basset	40	15	1	0	0	0	1	0	1	0
basset	26	23	1	0	0	0	1	0	1	0
basset	40	23	1	0	0	0	1	0	1	0
pointer	60	25	0	0	1	1	0	0	1	0
pointer	65	25	0	0	1	1	0	0	1	0
pointer	60	35	0	0	1	1	0	0	1	0
pointer	65	35	0	0	1	1	0	0	1	0
setter	53	20	0	0	1	1	0	0	1	0
setter	62	20	0	0	1	1	0	0	1	0
setter	53	32	0	0	1	1	0	0	1	0
setter	62	32	0	0	1	1	0	0	1	0
labrador	54	20	0	1	0	1	0	0	1	0
labrador	62	20	0	1	0	1	0	0	1	0
labrador	54	32	0	1	0	1	0	0	1	0
labrador	62	32	0	1	0	1	0	0	1	0
lévrier	55	25	0	0	1	1	0	0	1	0
lévrier	76	25	0	0	1	1	0	0	1	0
lévrier	55	48	0	0	1	1	0	0	1	0
lévrier	76	48	0	0	1	1	0	0	1	0
mastiff	75	100	1	0	0	0	1	0	0	1
ber-allema	58	26	0	0	1	0	1	0	0	1
ber-allema	65	26	0	0	1	0	1	0	0	1
ber-allema	58	34	0	0	1	0	1	0	0	1
ber-allema	65	34	0	0	1	0	1	0	0	1
dog-allema	76	50	0	0	1	0	1	0	0	1
dog-allema	80	50	0	0	1	0	1	0	0	1
dog-allema	76	70	0	0	1	0	1	0	0	1
dog-allema	80	70	0	0	1	0	1	0	0	1
doberman	68	20	0	0	1	0	1	0	0	1
doberman	70	20	0	0	1	0	1	0	0	1
doberman	68	26	0	0	1	0	1	0	0	1
doberman	70	26	0	0	1	0	1	0	0	1
saint-bern	70	55	1	0	0	0	1	0	0	1
saint-bern	70	80	1	0	0	0	1	0	0	1

TAB. 3.19: *Décomposition par sommets des 13 races de chiens*

Les variables obtenues suite à la décomposition sont de type numérique. On propose d'utiliser un découpage par histogramme puis un codage linéaire des variables numériques. L'étude des histogrammes des variables *Hauteur-Garrot* et *Poids* suggère un découpage du domaine de la variable *Hauteur* en trois modalités *H-*, *H+* et *H++* et un découpage du domaine de la variable *Poids* en trois modalités *P-*, *P+* et *P++*. On procède finalement au codage linéaire des variables *Hauteur* et *Poids* pour l'ensemble des sommets.

	Hauteur-Garrot			Poids		
Races	H-	H+	H++	P-	P+	P++
CAN	0.88	0.12	0.00	0.72	0.28	0.00
CAN	0.88	0.12	0.00	0.51	0.49	0.00
CAN	0.41	0.59	0.00	0.51	0.49	0.00
CAN	0.41	0.59	0.00	0.72	0.28	0.00
CHI	1.00	0.00	0.00	1.00	0.00	0.00
CHI	1.00	0.00	0.00	0.95	0.05	0.00
CHI	0.88	0.12	0.00	0.95	0.05	0.00
CHI	0.88	0.12	0.00	1.00	0.00	0.00
PEK	0.88	0.12	0.00	0.96	0.04	0.00
PEK	0.88	0.12	0.00	0.92	0.08	0.00
PEK	0.72	0.28	0.00	0.92	0.08	0.00
PEK	0.72	0.28	0.00	0.96	0.04	0.00
BAS	0.69	0.31	0.00	0.72	0.28	0.00
BAS	0.69	0.31	0.00	0.55	0.45	0.00
BAS	0.25	0.75	0.00	0.55	0.45	0.00
BAS	0.25	0.75	0.00	0.72	0.28	0.00
POI	0.00	0.62	0.38	0.51	0.49	0.00
POI	0.00	0.62	0.38	0.31	0.69	0.00
POI	0.00	0.47	0.53	0.31	0.69	0.00
POI	0.00	0.47	0.53	0.51	0.49	0.00
SET	0.00	0.84	0.16	0.61	0.39	0.00
SET	0.00	0.84	0.16	0.37	0.63	0.00
SET	0.00	0.56	0.44	0.37	0.63	0.00
SET	0.00	0.56	0.44	0.61	0.39	0.00
LAB	0.00	0.81	0.19	0.61	0.39	0.00
LAB	0.00	0.81	0.19	0.37	0.63	0.00
LAB	0.00	0.56	0.44	0.37	0.63	0.00
LAB	0.00	0.56	0.44	0.61	0.39	0.00
LEV	0.00	0.78	0.22	0.51	0.49	0.00
LEV	0.00	0.78	0.22	0.05	0.95	0.00
LEV	0.00	0.12	0.88	0.05	0.95	0.00
LEV	0.00	0.12	0.88	0.51	0.49	0.00
MAS	0.00	0.16	0.84	0.00	0.00	1.00
BER	0.00	0.69	0.31	0.49	0.51	0.00
BER	0.00	0.69	0.31	0.33	0.67	0.00
BER	0.00	0.47	0.53	0.33	0.67	0.00
BER	0.00	0.47	0.53	0.49	0.51	0.00
DOG	0.00	0.12	0.88	0.01	0.99	0.00
DOG	0.00	0.12	0.88	0.00	0.61	0.39
DOG	0.00	0.00	1.00	0.00	0.61	0.39
DOG	0.00	0.00	1.00	0.01	0.99	0.00
DOB	0.00	0.38	0.62	0.61	0.39	0.00
DOB	0.00	0.38	0.62	0.49	0.51	0.00
DOB	0.00	0.31	0.69	0.49	0.51	0.00
DOB	0.00	0.31	0.69	0.61	0.39	0.00
SBE	0.00	0.31	0.69	0.00	0.91	0.09
SBE	0.00	0.31	0.69	0.00	0.40	0.60

TAB. 3.20: *Codage par sommets des variables* Hauteur-Garrot *et* Poids

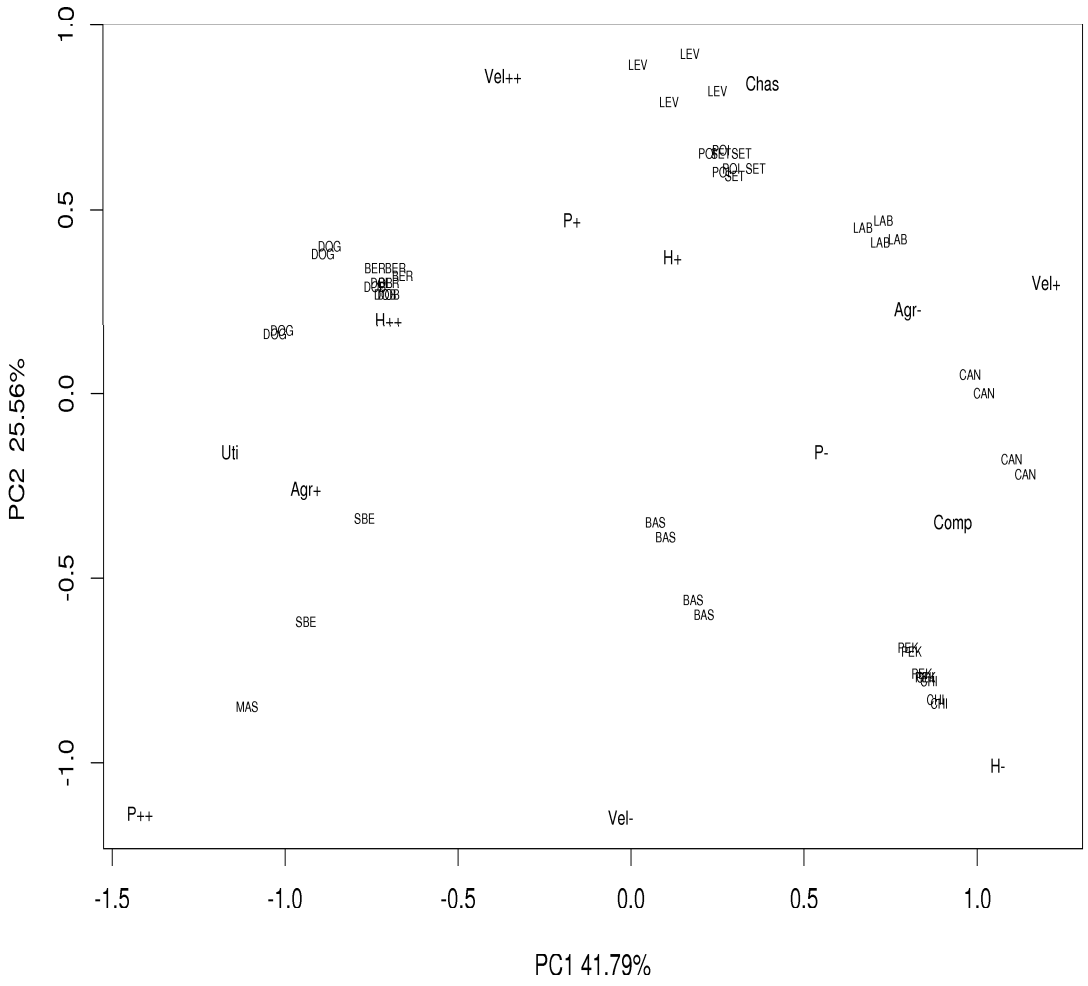
3.5.2 Résultats et interprétation

On applique l'Acm au tableau de donnée ainsi codé. Les trois premières valeurs propres $\lambda_1 = 0.55$ (41.79%), $\lambda_2 = 0.34$ (25.56%) et $\lambda_3 = 0.17$ (12.93%) reconstituent 80.28% de l'inertie totale. Les tableaux ci-dessous donnent les coordonnées factorielles des sommets associés à chaque objet chien et des modalités sur les 3 premiers axes d'inertie.

	Coordonnées factorielles des races de chiens		
	CP1	CP2	CP3
CAN	1.14	0.22	0.71
CAN	1.10	0.18	0.73
CAN	0.98	-0.04	0.82
CAN	1.02	0.00	0.80
CHI	0.89	0.83	-0.44
CHI	0.88	0.82	-0.43
CHI	0.85	0.77	-0.41
CHI	0.86	0.78	-0.42
PEK	0.85	0.77	-0.41
PEK	0.84	0.76	-0.41
PEK	0.80	0.68	-0.38
PEK	0.81	0.69	-0.38
BAS	0.21	0.59	-0.37
BAS	0.18	0.56	-0.35
BAS	0.07	0.35	-0.26
BAS	0.10	0.38	-0.29
POI	0.29	-0.62	-0.29
POI	0.25	-0.66	-0.27
POI	0.22	-0.65	-0.26
POI	0.26	-0.61	-0.29
SET	0.36	-0.61	-0.31
SET	0.31	-0.66	-0.28
SET	0.25	-0.64	-0.27
SET	0.30	-0.59	-0.30
LAB	0.77	-0.41	0.84
LAB	0.73	-0.46	0.87
LAB	0.67	-0.45	0.88
LAB	0.72	-0.40	0.85
LEV	0.25	-0.83	-0.35
LEV	0.16	-0.93	-0.29
LEV	0.02	-0.89	-0.28
LEV	0.11	-0.79	-0.33
MAS	-1.11	0.86	0.38
BER	-0.66	-0.31	-0.10
BER	-0.69	-0.35	-0.08
BER	-0.74	-0.33	-0.07
BER	-0.70	-0.30	-0.09
DOG	-0.87	-0.38	-0.03
DOG	-1.01	-0.17	0.11
DOG	-1.03	-0.16	0.11
DOG	-0.90	-0.38	-0.02
DOB	-0.70	-0.27	-0.11
DOB	-0.72	-0.29	-0.09
DOB	-0.74	-0.29	-0.09
DOB	-0.72	-0.26	-0.10
SBE	-0.77	0.34	0.06
SBE	-0.94	0.62	0.24

	Coordonnées factorielles des modalités		
	CP1	CP2	CP3
H-	1.06	1.00	-0.33
H+	0.12	-0.37	0.07
H++	-0.70	-0.20	0.12
P-	0.55	0.16	-0.23
P+	-0.18	-0.47	0.03
P++	-1.42	1.14	0.75
Vel-	-0.03	1.15	-0.29
Vel+	1.20	-0.29	1.96
Vel++	-0.37	-0.86	-0.42
Agr-	0.80	-0.23	-0.03
Agr+	-0.94	0.27	0.03
Comp	0.93	0.35	-0.03
Chas	0.38	-0.85	-0.26
Uti	-1.17	0.16	0.19

Le premier plan factoriel donnant la description simultanée des chiens et des modalités est :



Les paramètres d'aide à l'interprétation sont :

	Contributions absolues des modalités		
	CP1	CP2	CP3
H-	0.090	0.131	0.027
H+	0.002	0.031	0.002
H++	0.071	0.010	0.007
P-	0.048	0.007	0.026
P+	0.005	0.058	0.000
P++	0.086	0.091	0.081
Vel-	0.000	0.300	0.037
Vel+	0.080	0.008	0.690
Vel++	0.023	0.200	0.092
Agr-	0.126	0.016	0.000
Agr+	0.147	0.019	0.000
Comp	0.121	0.027	0.000
Chas	0.012	0.096	0.020
Uti	0.188	0.005	0.017

	Contributions relatives des modalités		
	CP1	CP2	CP3
H-	0.431	0.386	0.040
H+	0.032	0.299	0.009
H++	0.683	0.056	0.020
P-	0.574	0.050	0.097
P+	0.071	0.530	0.001
P++	0.341	0.220	0.099
Vel-	0.001	0.825	0.052
Vel+	0.262	0.016	0.698
Vel++	0.118	0.630	0.146
Agr-	0.755	0.060	0.001
Agr+	0.755	0.060	0.001
Comp	0.809	0.112	0.001
Chas	0.078	0.372	0.039
Uti	0.847	0.015	0.023

	Contributions absolues des chiens		
	CP1	CP2	CP3
CAN	0.045	0.003	0.056
CAN	0.042	0.002	0.060
CAN	0.033	0.000	0.076
CAN	0.036	0.000	0.072
CHI	0.028	0.040	0.022
CHI	0.027	0.039	0.021
CHI	0.025	0.033	0.019
CHI	0.026	0.034	0.020
PEK	0.025	0.034	0.019
PEK	0.025	0.033	0.019
PEK	0.022	0.027	0.016
PEK	0.023	0.027	0.016
BAS	0.002	0.020	0.015
BAS	0.001	0.018	0.014
BAS	0.000	0.007	0.008
BAS	0.000	0.008	0.009
POI	0.003	0.022	0.010
POI	0.002	0.025	0.008
POI	0.002	0.024	0.008
POI	0.002	0.021	0.009
SET	0.004	0.021	0.011
SET	0.003	0.025	0.009
SET	0.002	0.023	0.008
SET	0.003	0.020	0.010
LAB	0.021	0.010	0.079
LAB	0.018	0.012	0.085
LAB	0.016	0.011	0.086
LAB	0.018	0.009	0.081
LEV	0.002	0.039	0.014
LEV	0.001	0.049	0.010
LEV	0.000	0.045	0.009
LEV	0.000	0.036	0.013
MAS	0.043	0.042	0.016
BER	0.015	0.006	0.001
BER	0.016	0.007	0.001
BER	0.019	0.006	0.001
BER	0.017	0.005	0.001
DOG	0.027	0.008	0.000
DOG	0.035	0.002	0.001
DOG	0.037	0.001	0.001
DOG	0.028	0.008	0.000
DOB	0.017	0.004	0.001
DOB	0.018	0.005	0.001
DOB	0.019	0.005	0.001
DOB	0.018	0.004	0.001
SBE	0.021	0.007	0.000
SBE	0.031	0.022	0.006

	Contributions relatives des chiens		
	CP1	CP2	CP3
CAN	0.596	0.023	0.230
CAN	0.565	0.015	0.253
CAN	0.545	0.001	0.388
CAN	0.578	0.000	0.356
CHI	0.449	0.394	0.110
CHI	0.450	0.393	0.109
CHI	0.473	0.385	0.110
CHI	0.471	0.386	0.111
PEK	0.473	0.385	0.111
PEK	0.474	0.385	0.110
PEK	0.495	0.360	0.108
PEK	0.493	0.360	0.109
BAS	0.044	0.334	0.131
BAS	0.032	0.305	0.121
BAS	0.005	0.138	0.079
BAS	0.011	0.161	0.089
POI	0.130	0.583	0.130
POI	0.093	0.634	0.104
POI	0.072	0.632	0.103
POI	0.105	0.581	0.131
SET	0.168	0.480	0.124
SET	0.126	0.561	0.100
SET	0.096	0.630	0.113
SET	0.136	0.538	0.141
LAB	0.372	0.105	0.438
LAB	0.327	0.132	0.468
LAB	0.298	0.132	0.505
LAB	0.342	0.104	0.474
LEV	0.051	0.556	0.099
LEV	0.018	0.606	0.060
LEV	0.000	0.535	0.052
LEV	0.009	0.484	0.086
MAS	0.486	0.289	0.057
BER	0.471	0.107	0.011
BER	0.502	0.128	0.007
BER	0.601	0.124	0.006
BER	0.569	0.103	0.010
MAS	0.486	0.289	0.057
DOG	0.618	0.119	0.001
DOG	0.834	0.024	0.010
DOG	0.808	0.020	0.010
DOG	0.605	0.105	0.000
DOB	0.549	0.080	0.013
DOB	0.593	0.098	0.009
DOB	0.604	0.093	0.009
DOB	0.561	0.076	0.012
SBE	0.516	0.101	0.003
SBE	0.615	0.270	0.040

Discussion

Les modalités *H-, H+, H++*, par exemple, de la hauteur au garrot expriment, comme dans le cas classique, les régions de faibles, moyennes et fortes valeurs. Ainsi, la modalité *H-* caractérise essentiellement *Can*, *Pek* et *Chi*. Les degrés d'appartenance des sommets associés à *Can* varient dans [0.41 , 0.88] et ceux des sommets associés à *Chi* varient dans [0.88 , 1.00]. Dans le premier plan factoriel, on retrouve les principaux groupes dégagés dans le cas d'un codage linéaire croisé. Chaque objet chien incluant de la variation dans sa description est visualisé par un rectangle formé par l'ensemble de ses sommets. Remarquons que le chien *Mas* dont la hauteur au garrot et le poids sont numériques est représenté, comme dans le cas classique, par un point. La surface du rectangle associé à un objet est un indicateur de la variation intrinsèque à cet objet. La variation de la hauteur au garrot des races est restituée par la variation entre les sommets associés à chaque race. Dans le plan factoriel, les races présentant une grande variation sont *Can*, *Sbe*, *Lev* et *Bas* ; à l'opposé, *Chi*, *Pek* et *Poi* sont caractérisés par une faible variation.

3.6 Application du codage sans décomposition en ACM

3.6.1 Découpage des variables intervalles par histogramme

On procède à la construction des histogrammes des variables intervalles *Hauteur-Garrot* et *Poids*. Pour cela, on discrétise le domaine de la variable intervalle en sous-intervalles d'amplitude égale à la plus petite amplitude non nulle observée. L'histogramme de la variable est obtenu en calculant pour chaque sous-intervalle, issues de la discrétisation, sa densité de recouvrement par l'ensemble des intervalles observés pour la variable en question. Les histogrammes des variables *Hauteur-Garrot* et *Poids* ainsi définis figurent en 3.17 et 3.18.

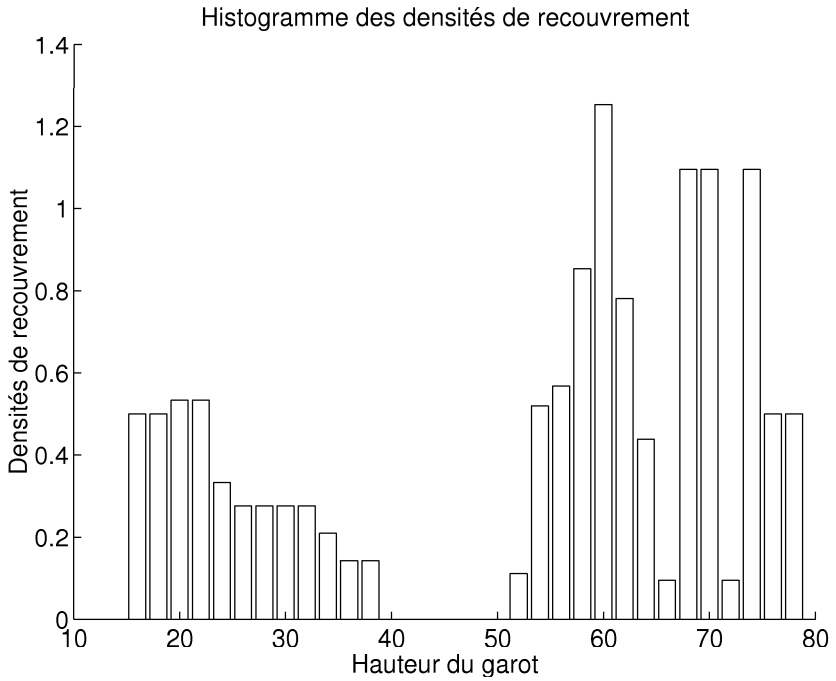


FIG. 3.17:

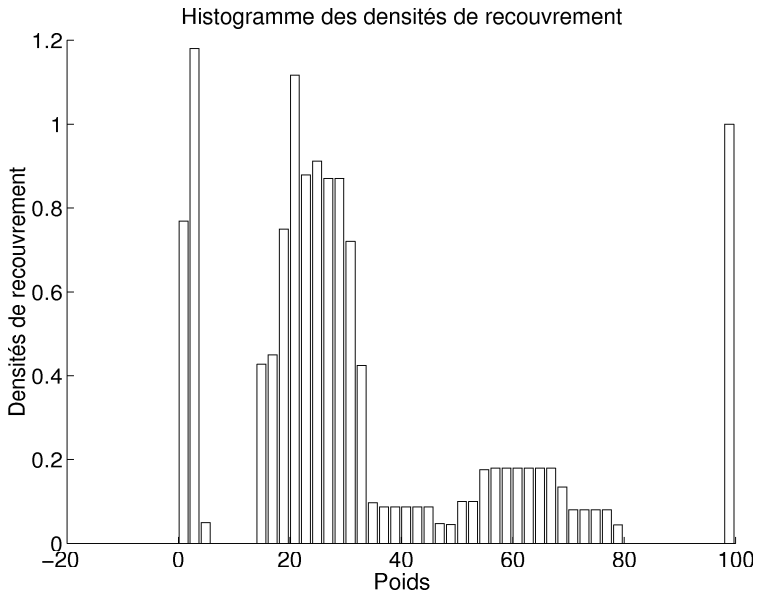


FIG. 3.18:

L'étude des histogrammes des variables suggère un découpage de la variable *Hauteur-Garrot* en deux modalités notées *H-* , *H+* définies par les sous intervalles $[16, 45]$ et $[45, 80]$ et un découpage de la variables *Poids* en trois modalités *P-*, *P+* et *P++* définies par les intervalles $[0.9, 30]$, $[30, 50]$ et $[50, 100]$.

3.6.2 Codage basé sur la distance de MOORE

Le codage des variables intervalles d'après la fonction d'appartenance définie en 3.12 est donnée par le tableau suivant :

	Hauteur-Garrot		Poids		
Races	H-	H+	P-	P+	P++
CAN	0.7655	0.2345	0.4486	0.3887	0.1626
CHI	0.9403	0.0597	0.5439	0.3812	0.0749
PEK	0.8443	0.1557	0.5334	0.3816	0.0850
BAS	0.6983	0.3017	0.4536	0.3885	0.1579
POI	0.3046	0.6954	0.3809	0.4156	0.2036
SET	0.3661	0.6339	0.4087	0.4032	0.1881
LAB	0.3607	0.6393	0.4087	0.4032	0.1881
LEV	0.2508	0.7492	0.3369	0.4281	0.2350
MAS	0.0773	0.9227	0	0.3311	0.6689
BER	0.3162	0.6838	0.3808	0.4163	0.2029
DOG	0	1.0000	0.2393	0.3839	0.3768
DOB	0.2095	0.7905	0.4267	0.3995	0.1738
SBE	0.1897	0.8103	0.2066	0.3567	0.4366

3.6.3 Résultats et interprétation

On applique l'Acm au tableau de donnée ainsi codé. Les trois premières valeurs propres $\lambda_1 = 0.47$ (43.06%), $\lambda_2 = 0.30$ (27.31%) et $\lambda_3 = 0.16$ (14.57%) reconstituent 84.94% de l'inertie totale. Les coordonnées factorielles des chiens et des modalités, sur les 3 premiers axes d'inertie, sont données par les tableaux suivants :

	Coordonnées factorielles des races de chiens		
	CP1	CP2	CP3
CAN	1.12	-0.25	0.82
CHI	0.79	-0.76	-0.38
PEK	0.77	-0.74	-0.36
BAS	-0.08	-0.20	-0.46
POI	0.19	0.75	-0.35
SET	0.21	0.74	-0.36
LAB	0.73	0.56	0.62
LEV	0.17	0.76	-0.34
MAS	-0.78	-0.56	0.09
BER	-0.76	0.07	0.20
DOG	-0.87	0.12	0.26
DOB	-0.78	0.09	0.21
SBE	-0.70	-0.57	0.05

	Coordonnées factorielles des modalités		
	CP1	CP2	CP3
H-	0.56	-0.31	-0.15
H+	-0.3	8 0.22	0.11
P-	0.26	0.01	-0.07
P+	0.01	0.05	-0.01
P++	-0.41	-0.10	0.12
Vel-	-0.00	-1.03	-0.53
Vel+	1.35	0.28	1.81
Vel++	-0.45	0.77	-0.16
Agr-	0.83	0.28	-0.12
Agr+	-0.96	-0.32	0.14
Comp	1.30	-1.07	0.06
Chas	0.35	0.95	-0.45
Uti	-1.13	-0.31	0.40

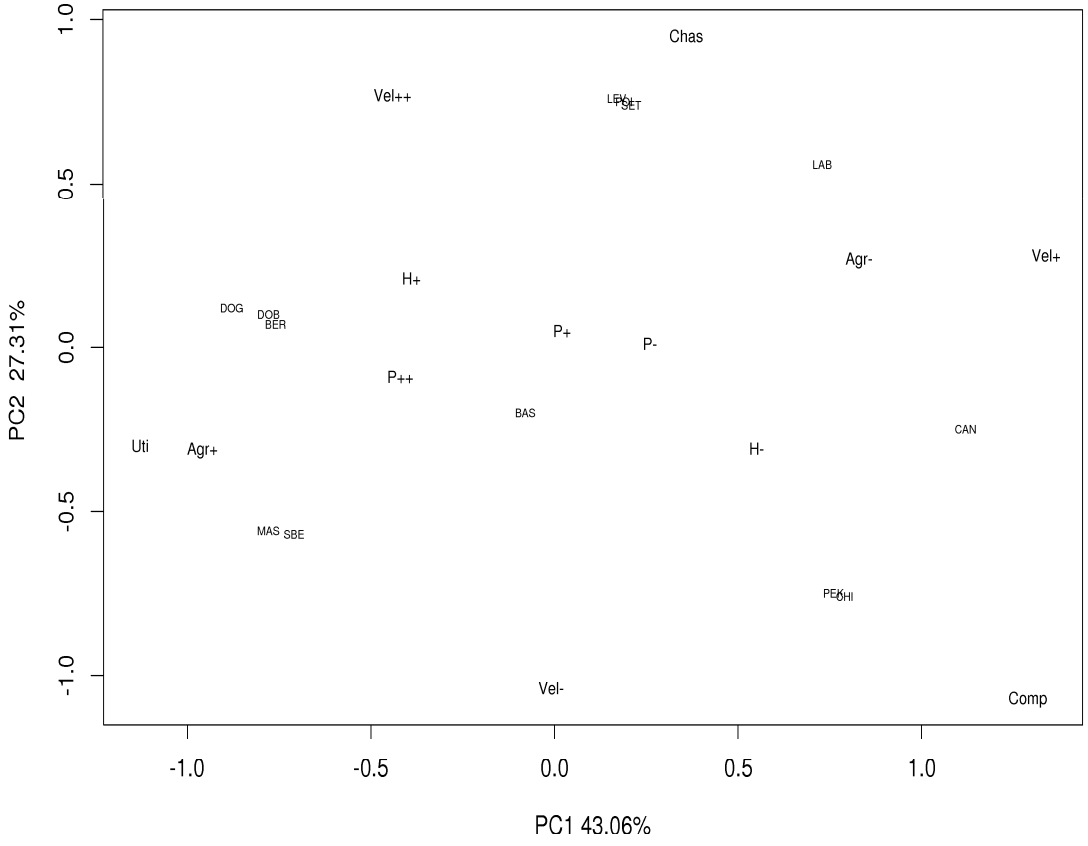
Paramètres d'aide à l'interprétation

	Contributions absolues des modalités		
	CP1	CP2	CP3
H-	0.05	0.03	0.01
H+	0.04	0.02	0.01
P-	0.01	0.00	0.00
P+	0.00	0.00	0.00
P++	0.02	0.00	0.00
Vel-	0.00	0.27	0.14
Vel+	0.12	0.01	0.63
Vel++	0.04	0.18	0.01
Agr-	0.15	0.03	0.01
Agr+	0.18	0.03	0.01
Comp	0.16	0.18	0.00
Chas	0.02	0.23	0.09
Uti	0.21	0.02	0.08

	Contributions relatives des modalités		
	CP1	CP2	CP3
H-	0.610	0.190	0.046
H+	0.612	0.190	0.045
P-	0.444	0.001	0.035
P+	0.060	0.568	0.011
P++	0.415	0.022	0.034
Vel-	0.000	0.672	0.179
Vel+	0.327	0.014	0.597
Vel++	0.168	0.510	0.021
Agr-	0.796	0.085	0.017
Agr+	0.797	0.085	0.017
Comp	0.501	0.346	0.001
Chas	0.079	0.564	0.122
Uti	0.801	0.058	0.099

	Contributions absolues des chiens		
	CP1	CP2	CP3
CAN	0.204	0.016	0.325
CHI	0.102	0.148	0.069
PEK	0.095	0.141	0.064
BAS	0.001	0.010	0.103
POI	0.006	0.144	0.059
SET	0.007	0.140	0.062
LAB	0.086	0.081	0.187
LEV	0.005	0.148	0.055
MAS	0.098	0.080	0.004
BER	0.093	0.001	0.019
DOG	0.124	0.004	0.032
DOB	0.098	0.002	0.021
SBE	0.080	0.084	0.001

	Contributions relatives des chiens		
	CP1	CP2	CP3
CAN	0.612	0.031	0.330
CHI	0.441	0.405	0.100
PEK	0.435	0.409	0.099
BAS	0.007	0.042	0.226
POI	0.048	0.764	0.165
SET	0.060	0.747	0.176
LAB	0.332	0.198	0.244
LEV	0.037	0.771	0.153
MAS	0.509	0.264	0.006
BER	0.719	0.006	0.051
DOG	0.806	0.016	0.070
DOB	0.735	0.010	0.054
SBE	0.517	0.342	0.003



Discussion

Les modalités issues d'un codage sans décomposition expriment, comme dans le cas classique, des régions de valeurs faibles, moyennes, fortes, etc. Les résultats du codage de la variable hauteur au garrot et poids sont similaires à ceux obtenus dans le cas du codage par sommets. Par exemple, la modalité *H-* caractérise essentiellement, *Chi* (0.94), *Pek* (0.84) , *Can* (0.76) et *Bas* (0.69). On distingue dans le premier plan factoriel quatre groupes similaires à ceux dégagés dans le cas d'un codage linéaire croisé ou par sommets : {*Mas, Sbe, Dog, Dob, Ber*}, {*Lev, Poi, Set, Lab*}, {*Chi, Pek, Can*} et {*Bas*}. Cependant, on constate que les groupes identifiés sont beaucoups plus compactes du faite de la normalisation introduite dans la fonction d'appartenance. Par ailleur, on constate une séparation plus distinguée entre les deux sous-groupes {*Mas, Sbe*} et {*Dog, Dob,Ber*}.

Conclusion

Résultats et persepectives

Dans le premier chapitre on s'intéresse à l'extension de l'analyse en composantes principales à des tableaux d'intervalles. On propose deux approches : *la méthode des sommets* et *la méthode des centres.*

• *La méthode des centres* La variance prise en compte dans la méthode des centres est la variance des centres des hyper-rectangles. La variation autour des centres est restituée, dans les plans factoriels, après l'analyse. La méthode des centres est une *analyse inter hyper-rectangles.*

• *La méthode des sommets* La variance prise en compte dans la méthode des sommets est la variance des sommets des hyper-rectangles. Elle se décompose en la somme de deux effets : un effet fixe, exprimant la variance des centres des hyper-rectangles et un effet variable exprimant la variance tenant compte des amplitudes des intervalles. La méthode des sommets est une *analyse inter et intra hyper-rectangles.*

• *Distributions au sein des intervalles* En considérant des distributions usuelles symétriques au sein des intervalles (Normale, Uniforme, Triangulaire, etc.), on montre que la variance prise en compte se décompose également en la somme de deux effets : un effet fixe correspondant à la variance des centres des hyper-rectangles et un effet variable tenant compte de la distribution autour des centres. Dans le cas d'une pondération équirépartie des sommets, la variance dans la méthode des sommets majore celles obtenues en considérant une distribution symétrique au sein des intervalles. La variance dans la méthode des centres minore toutes les autres variances.

• *Intervalles munis de contraintes de domaines* Dans le cas de données intervalles munies de contraintes de domaines, un hyper-rectangle ne constitue plus un volume homogène, mais peut comporter une ou plusieurs régions non valides. Pour tenir compte des contraintes de domaines dans l'analyse on propose, selon la stratégie de pondération adoptée proportionnelle aux volumes ou inversement proportionnelle aux volumes, de décroître, respectivement, de croître le poids d'un objet H_i proportionnellement aux volumes des hyper-rectangles contraintes qui lui sont associés.

• *Visualisation à différents niveaux de qualité de représentation* On propose une procédure itérative, qui intègre les qualités de représentation (contributions relatives) des sommets lors de la construction et de la visualisation des objets. Pour chaque couple de composantes principales, elle fournit une succession de plans factoriels, chacun associé à une seuil de qualité de représentation. Un objet est représenté, dans les plans factoriels, par un rectangle, un segment ou un point.

Dans le second chapitre on s'intéresse à l'extension de l'analyse des correspondances multiples à des variables de type intervalle. On propose trois différentes techniques de codage flou : *le codage croisé*, *le codage par sommets* et *le codage sans décomposition.*

Le principe des deux premiers codage est la transformation des variables intervalles en variables numériques. Le codage flou des variables intervalles est déduit à partir du codage flou des variables numériques issues de la transformation.

• *Le codage croisé* Il assure le codage flou d'une variable de type intervalle en une variable qualitative à k modalités. Il se base sur la décomposition de chaque variable intervalle en deux variables numériques. Le découpage du domaine de la variable intervalle est défini par le croisement des découpages associés aux variables numériques. De même, les fonctions d'appartenance associées aux classes de la variable intervalle sont définies par le croisement des fonctions d'appartenance associées aux classes des variables numériques. Les modalités issues du codage croisé ne représentent pas uniquement des régions de valeurs (comme dans le cas numérique) mais également des niveaux de variation. Une modalité représente une classe d'intervalles, proches par leur po-

sitionnement et leur amplitude de variation. Ainsi, l'information de variation intrinsèque à un objet est déduite à partir des modalités qui le caractérisent.

• *Le codage par sommets* De manière similaire, la première étape du codage par sommets est la transformation des variables intervalles en des variables numériques. Pour cela, la description d'un objet par des données intervalles, est substituée par les descriptions numériques de ses sommets. Le codage flou des objets sont déduits à partir des codages flous de leurs sommets. Remarquons, que les modalités issues du codage par sommets ne représente que des régions de valeurs, comme dans le cas classique. La variation inhérente à un objet est déduite à partir de la variation des sommets qui le composent.

• *Le codage sans décomposition* Le troisième codage, comme son nom l'indique, ne se base pas sur la décomposition des variables intervalles en des variables numériques mais sur l'extension des outils de codage des variables numériques à des variables intervalles.

a) Découpage en classes On s'intéresse, dans une première partie, aux techniques de découpage d'une variable en classes. On propose, une extension du découpage en classes, d'effectifs égaux et à partir d'un histogramme, à des variables intervalles. Pour cela, on propose une généralisation, respectivement, de la notion de fonction de répartition et d'histogramme à une distribution d'intervalles.

b) Fonction d'appartenance Dans une deuxième partie, on s'intéresse aux fonctions d'appartenance qui mesurent le degré d'appartenance d'un intervalle à une classe d'intervalles. Après l'étude et la définition des propriétés que doit vérifier une fonction d'appartenance associée à une classe d'intervalles, on propose une fonction d'appartenance basée sur la distance de Moore.

Les perspectives, à court terme, de ce travail de thèse portent sur les points suivants :

• *La pondération des sommets* Dans la technique de pondération des sommets, le poids d'un sommet est défini par le produit des poids de ses coordonnées, sous l'hypothèse d'indépendance entre les distributions à l'intérieur des intervalles. Plus généralement, on peut envisager d'étudier les propriétés ainsi que les interprétations que cela engendre dans le cas où l'on ne fait plus

cette hypothèse.

• *La méthode itérative de visualisation* Cette méthode bien qu'elle permette une visualisation plus claire, elle demeure néanmoins coûteuse quand le nombre de variables de type intervalle est important. En effet, cette technique implique le traitement de tous les sommets pour sélectionner ceux ayant une contribution relative supérieure à un seuil donné. Pour réduire la complexité, l'objectif consiste à restituer la variation autour de chaque centre et à pouvoir mesurer la contribution relative de celle-ci, sans avoir à traiter les sommets concernés.

• *Robustesse de la méthode des sommets* Dans le cas où les bornes des intervalles correspondent à des valeurs atypiques des distributions à l'intérieur des intervalles, les résultats de l'analyse (i.e axes factoriels, valeurs propres, etc.) risquent d'être perturbés par de tels éléments. Deux cas sont alors à envisager. Dans un premier cas, si l'on dispose des distributions à l'intérieur des intervalles, le problème des valeurs atypiques est soulevé en définissant les intervalles non pas à partir des valeurs extrêmes mais de manière plus précise, par exemple, en prenant les intervalles de confiances à un ou deux écart-types robustes autour de la moyenne. Dans le cas où les distributions ne sont pas connues, des simulations devront être effectuer afin d'étudier la validité des résultats.

• *Le choix du seuil* α Dans la technique itérative de visualisation les seuils α sont fixés a priori à 0.2, 0.5 et 0.6. Si la part de l'inertie totale restituée par le plan factoriel est très élevée, alors il est inutile de fixer des seuils α faibles ; les plans factoriels associés risquent d'être identiques. Pour cela, on peut envisager de fixer les seuils α ainsi que le nombre d'itérations à effectuer en fonction du pourcentage de l'inertie totale restitué dans les plans factoriels de niveau zéro.

• *Le codage sans décomposition* La normalisation utilisée dans la fonction d'appartenance du codage sans décomposition risque d'atténuer l'effet du codage. Notre objectif est de définir une fonction d'appartenance introduisant la contrainte de normalisation clors de la mesure de distance ou de dissimilarité effectuée entre les intervalles et les modalités.

Bibliographie

[Anwar93] Anwar (N.). – *Micro-agrregation - The small Aggregates Method.* – Rapport technique, Internal Report, Luxembourg, Eurostat, 1993.

[Bandemer et al.92] Bandemer (H) et Näther (W.). – *Fuzzy Data Analysis.* – Kluwer Academic, 1992, *B.*

[Benzecri73] Benzécri (J.P.). – *L'analyse des données, Tome 2 : L'analyse des correspondances.* – Dunod, Paris, 1973.

[Bordet73] Bordet (J.P). – *Études de données géophysiques.* – Thèse de PhD, Université Paris VI, 1973. Thèse de 3ème cycle.

[Brito91] Brito (P.). – *Analyse de données symboliques. Pyramides d'héritage.* – Thèse, Université Paris ix-Dauphine, 1991.

[Calahan72] Calahan (D.). – *Computer-Aided Network Design*, chap. Tolerence Analysis, p. 165. – McGraw-Hill, 1972.

[Carvalho92] Carvalho (F.A.T De). – *Méthode descriptives en Analyse des Données Symboliques.* – Thèse de PhD, Paris IX Dauphine, 1992.

[Carvalho95] Carvalho (F.A.T De). – Histograms in symbolic data analysis. *Annals of Operations Research*, vol. 55, 1995, pp. 299–322.

[Caussinus92] Caussinus (H.). – *Modèles pour l'analyse des données multidimentionnelles*, chap. 3, pp. 61–81. – Economica, 1992, journées d'étude en statistiques édition.

[Cazes et al.97] CAZES (P.), CHOUAKRIA (A.), DIDAY (E.) et SCHEKTMAN (Y.). – Extension de l'analyse en composantes principales à des données intervalles. *Revue de statistique appliquée*, vol. XLV, n° 3, 1997, pp. 5–24.

[Cazes90] CAZES (P.). – Codage d'une variable continue en vue de l'analyse des correspondances. *Revue de statistique appliquée*, vol. XXXVIII, n° 3, 1990, pp. 35–51.

[Chavent97] CHAVENT (M.). – *Analyse des données symboliques. Une méthode divisive de classification.* – Thèse de PhD, Université Paris IX-Dauphine, décembre 1997.

[Chouakria et al.95a] CHOUAKRIA (A.), CAZES (P.) et DIDAY (E.). – Extension de l'analyse factorielle des correspondances multiples à des données de type intervalle et de type ensemble. *SFC'95 : Actes de la* 3^{eme} *rencontre de la Société Francophone de Classification.* – Namur, septembre 1995.

[Chouakria et al.95b] CHOUAKRIA (A.), DIDAY (E.) et CAZES (P.). – Extension of the principal component analysis to interval data. *NTTS'95 : New Techniques and Technologies for Statistics,*. – Bonn, November 1995.

[Chouakria et al.96] CHOUAKRIA (A.), VERDE (R.), DIDAY (E.) et CAZES (P.). – Généralisation de l'analyse factorielle des correspondances multiples à des objets symboliques. *SFC'96 : Actes des la* 4^{eme} *rencontre de la la Société francophone de classification.* – Vannes, septembre 1996.

[Chouakria et al.97] CHOUAKRIA (A.), DIDAY (E.) et CAZES (P.). – Généralisation, en vue d'une ACM, du découpage en classes d'effectifs égaux à des variables de type intervalle. ASU *'97 Association de la Statistique et de ses Utilisateurs.* – 1997.

[Chouakria et al.98a] CHOUAKRIA (A.), DIDAY (E.) et CAZES (P.). – An improved factorial representation of symbolic objects. *KESDA'98 : Knowledge Extraction from Statistical Data.* EUROSTAT. – Luxembourg, April 1998.

[Chouakria et al.98b] CHOUAKRIA (A.), DIDAY (E.) et CAZES (P.). – Vertices principal components analysis with an improved factorial representation. *Advances in Data Science and Classification*, éd. par RIZZI (A.), VICHI (M.) et Bock (H.). pp. 397–402. – Rome, July 1998. ISBN 3-540-64641-8.

[Chouakria94] CHOUAKRIA (A.). – *Extension de méthodes de réduction de dimension à des données symboliques.* – Mémoire de D.E.A à l'université Paris-IX Dauphine, septembre 1994.

[Defays et al.93] DEFAYS (D.) et NANOPOULOS (P.). – The small aggregates method. *The 92 Symposium on "Design and Analysis of Longitudinal Surveys", Ottawa, Statistics Canada.* – 1993.

[Diday et al.96] DIDAY (E.), ÉMILION (R.) et HILLALI (Y.). – Symbolic data analysis of probabilistic objects by capacities and credibilities. XXXVIII *Societa'Italiana Di Statistica.* – Rimini, April 1996.

[Diday89a] DIDAY (E.) (édité par). – *Data Analysis, Learning Symbolic and Numeric Knowledge.* – Nova Science, September 1989.

[Diday89b] DIDAY (E.). – Introduction à l'approche symbolique en analyse des données. RAIRO, vol. 23, n° 2, 1989.

[Diday91] DIDAY (E.). – *Des objets de l'analyse des données à ceux de l'analyse des connaissances. Induction symbolique et numérique à partir de données.* – E. DIDAY and Y. KODRATOFF Cépaduès, 1991.

[Diday95] DIDAY (E.). – Probabilistic objects for a symbolic data analysis. *Discrete Mathematics and Theorical Computers*, vol. 19, 1995.

[Droesbecke et al.92] DROESBECKE (JJ.), FICHET (B.) et TASSI (P.). – *Modèles pour l'analyse des données multidimentionnelles.* – Economica, 1992, journées d'étude en statistiques édition.

[Dubois et al.85] DUBOIS (D.) et PRADE (H.). – *Théorie des possibilités : Application à la représentation des connaissances en informatique.* – Masson, 1985.

[Eurostat97] Eurostat (Statistiques En Bref.). – *Le Chômage dans les régions de l'Union Européenne en 1996.* – Rapport technique n° 3, Eurostat ISSN 1025-0417, 1997.

[Fisher58] FISHER (R.A.). – On grouping for maximum homogeneity. *Journal of the American Statistical Association*, vol. 53, 1958, pp. 789–798.

[Gallego82] GALLEGO (F . J). – Codage flou en analyse des correspondances. *Les cahiers de l'analyse des données*, vol. VII, n° 4, 1982, pp. 413–430.

[Gettlersumma92] GETTLER-SUMMA (M.). – Factorial axis interpretation by symbolic objects. *Journées Symboliques - Numériques.* – LISE CEREMADE, Université Paris IX-Dauphine, 1992.

[Ghermani et al.77] GHERMANI (B.M), ROUX (C.) et ROUX (M.). – Sur le codage logique des données héterogènes. *Les cahiers de l'analyse des données*, vol. 1, 1977, pp. 115–118.

[Gowda et al.91a] GOWDA (K. Chidananda) et DIDAY (E.). – Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, vol. 24, n° 6, 1991.

[Gowda et al.91b] GOWDA (K. Chidananda) et DIDAY (E.). – Unsupervised learning through symbolic clustering. *Pattern Recongnition Letters*, vol. 12, 1991.

[Gowda et al.92] GOWDA (K. Chidananda) et DIDAY (E.). – Symbolic clustering using a new similarity measure. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, n° 2, March/April 1992.

[Grabisch et al.95] GRABISCH (M.), HUNG (T.), ELBERT (A.) et WALKER (A.). – *Fundamentals of uncertainty calculi with applications to fuzzy inference.* – Series B: Mathematical and statistical methods, 1995.

[Guitonneau et al.77] Guitonneau (G.G) et M.Roux. – Sur la taxinomie du genre erodieum. *Les cahiers de l'analyse des données*, vol. 1, 1977, pp. 97–113.

[Hillali98] Hillali (Y.). – *Analyse et modélisation des données probabilistes : Capacités et lois multidimensionnelles.* – Thèse de PhD, Université Paris IX-Dauphine, mars 1998.

[Hotelling33] Hotelling (H.). – Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, vol. 24, 1933.

[Ichino et al.94] Ichino (M.) et Yaguchi (H.). – Generalized minkowski metrics for mixed feature-type data analysis. *IEEE Transaction On Systems, Man and Cybernetics*, vol. 24, n° 4, April 1994.

[Kanade73] Kanade (T.). – *Picture processing by computer complex and recognition of human faces.* – Rapport technique, Kyoto University, Department of Information Science, November 1973.

[Kirby et al.90] Kirby (M.) et Sirovich (L.). – Application of the karhunen-loève procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, n° 1, January 1990, pp. 103–108.

[Lafayedemicheaux78] Lafaye de Micheaux (D.). – *Approximations d'analyses canoniques nonlinéaires de variables aléatoires et analyses factorielles privilégiantes.* – Thèse de PhD, Université de Nice, 1978. Thèse de Docteur-Ingénieur.

[Lebart et al.95] Lebart (L.), Morineau (A.) et Piron (M.). – *Statistique exploratoire multidimentionnelle.* – Dunod, 1995.

[Lebbe et al.91] Lebbe (J.) et Vignes (R.). – Génération de graphes d'identification à partir de descriptions de concepts. *Induction symbolique et numérique à partir de données*, éd. par Kodratoff (Y.) et Diday (E.). – Cépaduès, 1991.

[Lefoll79] Le Foll (Y.). – *Sur les propriétés de l'analyse des correspondances pour diverses formes complètes de données.* – Thèse de PhD, Université de Paris VI, 1979. Thèse de 3ème cycle.

[Leroy et al.96] Leroy (B.), Chouakria (A.), herlin (I.) et Diday (E.). – Approche géométrique et classification pour la reconnaissance de visage. *RFIA'96, Reconnaissance des Formes et Intelligence Artificielle.* Inria and Irisa and Cnrs, pp. 548–557. – janvier 1996.

[Martin80] Martin (J.F). – *Codage flou et ses applications en statistique.* – Thèse de PhD, Université de Pau et des Pays de L'adour, 1980. Thèse de 3ème cycle.

[Mfoumoune98] Mfoumoune (E.). – *Les aspects algorithmiques de la classification ascendante pyramidale et incrémentale.* – Thèse de PhD, Université Paris-XI Dauphine, février 1998.

[Moore66] Moore (R. E.). – *Interval Analysis.* – Prentice Hall, Englewod Cliffs, New Jersey, 1966.

[Nagabhushan97] Nagabhushan (P.). – Dimensionality reduction of symbolic data. *Indo-French Workshop on Symbolic Data Analysis and its Applications.* Lise-Ceremade, Paris IX Dauphine. – 1997.

[Nagaraj93] Nagaraj (N.S.). – *Dimensionality reduction of Symbolic Data.* – Thèse de PhD, sri jayachamarajendra College of Engineering University of Mysore, 1993.

[Perinel96] Perinel (E.). – *Segmentation et analyse des données symboliques. Application à des données probabilistes imprécises.* – Thèse de PhD, Université Paris IX-Dauphine, 1996.

[Polaillon et al.96] Polaillon (G.) et Diday (E.). – *Galois Lattices construction and application in Symbolic Data Analysis.* – Rapport technique, Cahiers de Mathématiques de la décision N^o 9631, Ceremade, Université Paris IX Dauphine, 1996.

[Rijckevorsel et al.88] Rijckevorsel (J.L.A) et De Leeuw (J.). – *Component and correspondence analysis. Dimension reduction by functional approximation.* – Wiley Sons, 1988.

[Rijckevorsel87] Rijckevorsel (J.L.A). – *The application of fuzzy coding and horseshoes in multiple correspondence analysis.* – DSWO Press, Leiden, 1987.

[Rijckevorsel88] Rijckevorsel (J.L.A). – *Fuzzy coding and B-splines*, chap. 2, pp. 33–54. – Wiley, 1988.

[Saporta90] Saporta (G.). – *Probabilités, analyse des données et statistique.* – Technip, 1990.

[Schweitzer et al.61] Schweitzer (B.) et Sklar (A.). – Asscociative functions and statistical triangle inequalities. *Publ. Math. Debrecen*, vol. 8, 1961, pp. 169–186.

[Stephan98] Stéphan (V.). – *Construction d'objets symboliques par synthèse des résultats de requêtes* Sql. – Thèse de PhD, Université Paris IX-Dauphine, janvier 1998.

[Turk et al.91] Turk (M.) et Pentland (A.). – Eigenfaces for recognition. *Cognitive Neuroscience*, vol. 3, n° 1, 1991, pp. 71–86.

[Vignes91] Vignes (R.). – *Caractérisation automatique de groupes biologiques.* – Thèse, Université Pierre et Marie Curie – Paris vi, avril 1991. Sciences de la vie.

[Ziani96] Ziani (D.). – *Sélection de variables sur un ensemble d'objets symboliques.* – Thèse, Université Paris ix–Dauphine, 1996.